What will be left for us to work on?

normaltech.ai/p/what-will-be-left-for-us-to-work

Arvind Narayanan July 13, 2026

AI AS NORMAL TECHNOLOGY

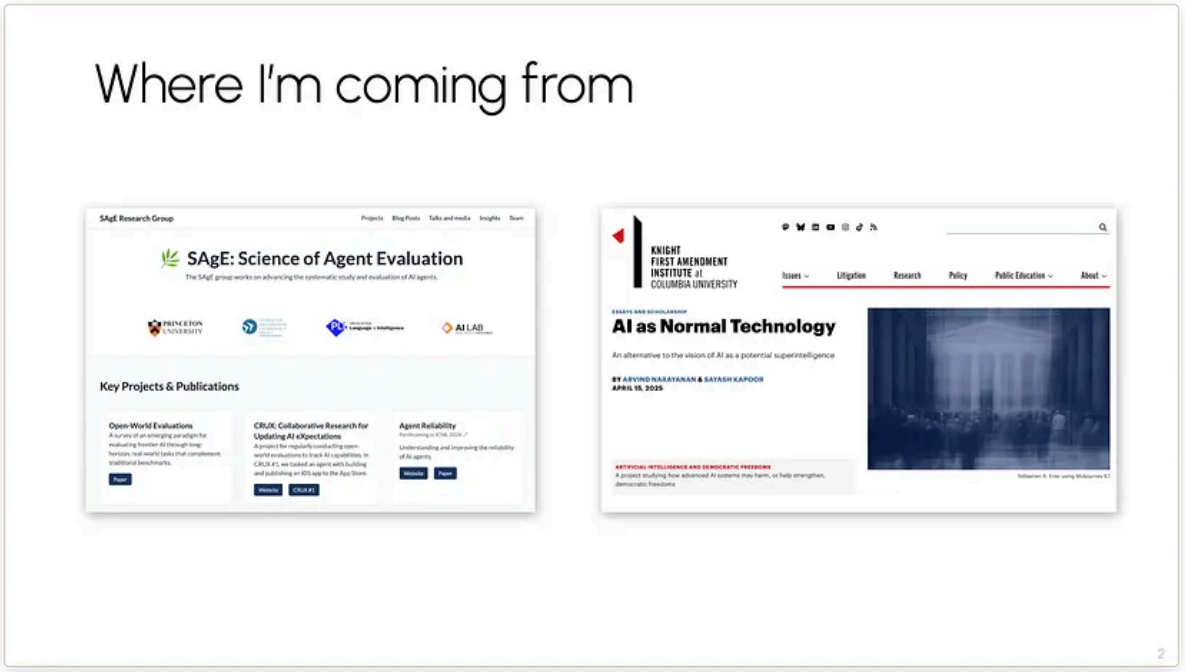
I had the honor of giving a keynote at the International Conference on Machine Learning in Seoul last week titled “What will be left for us to work on?” I addressed the widespread anxiety about how we should adapt as AI capabilities increase. I was thrilled by the talk’s reception, so I have made my slides available **here**, annotated with a lightly edited transcript. You can also view them below right here on this page, but the **online version** has animations, clickable links, and a much nicer experience overall.

I made three arguments. First, the *AI as Normal Technology* framework is a correct and useful as a way to think about AI’s impacts, unless and until there is some future discontinuity such as through recursive self-improvement. Second, even though we should take recursive self-improvement seriously, there is no milestone that companies might achieve in the lab that will suddenly put us all out of work. Third and finally, jobs of the future will be radically different, and a lot of adaptation will be needed. I shared my thinking about what this might look like and ended with a vision of human/AI “co-superintelligence”.

What will be left for us to work on?
Arvind Narayanan
Professor of Computer Science
Co-author of AI as Normal Technology
PRINCETON UNIVERSITY
https://www.normaltech.ai/
Essays related to this talk

Now is a time of great excitement in AI, but it’s also a time of great anxiety in the AI community. I want to address that anxiety head on. How do we prepare for a future

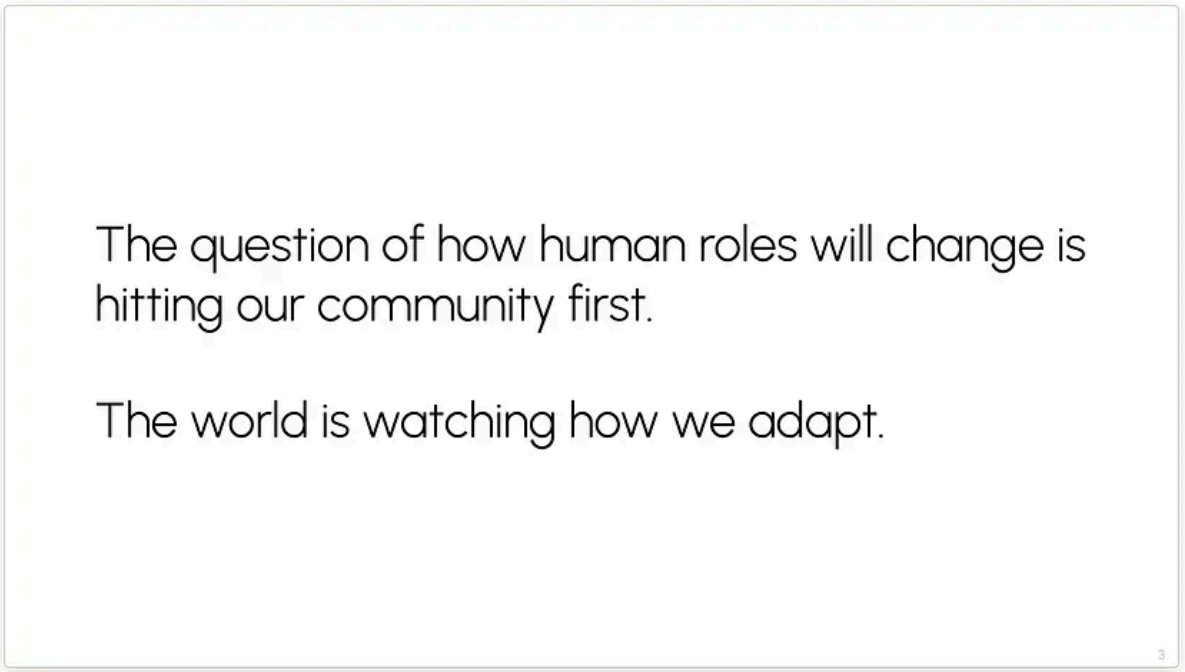
where AI will become capable of doing more and more of the work that we do today?



I lead a team at Princeton University trying to advance the science of AI agent evaluation. We try to go beyond the usual claims of “Look, capability is going up on benchmarks!” Those claims tend to be misinterpreted by the broader public as implying that agents are soon about to take all our jobs.

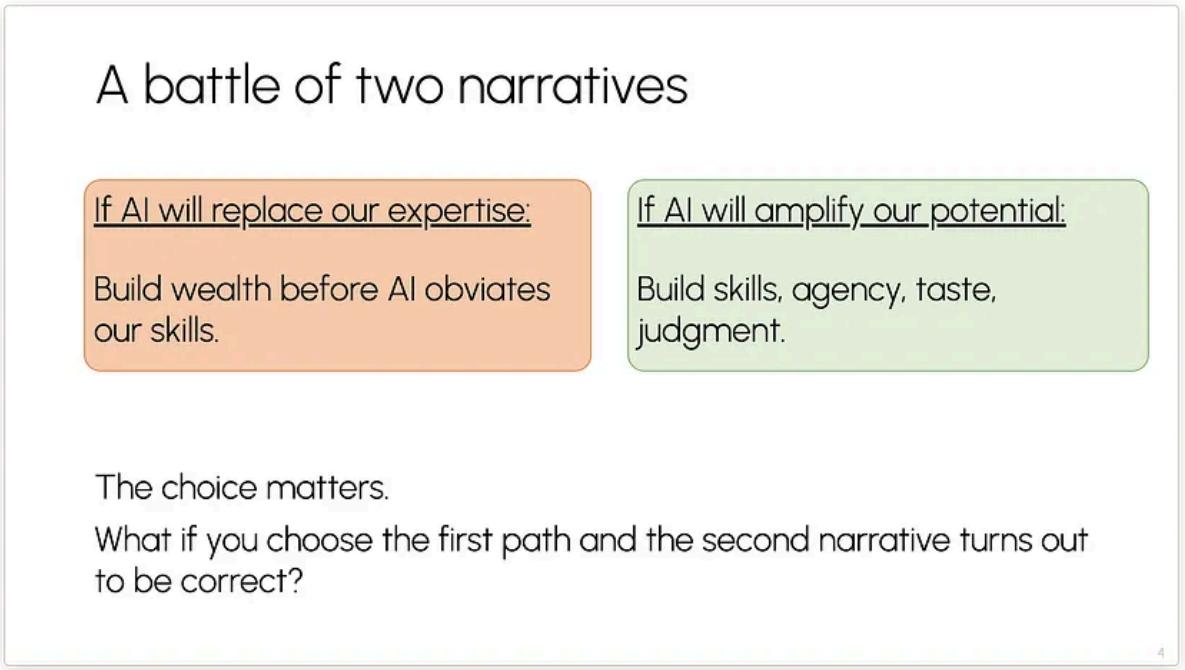
Maybe that will happen. But in our work we try to understand the factors beyond capability that matter for real-world deployment, and bring that understanding into evaluations.

The work that I’m better known for is the essay I co-authored with Sayash Kapoor called AI as Normal Technology. It’s a way to think about the medium-term future of AI and how to adapt to it — and in turn how to adapt it to the needs of society and the economy.



So we've been going around writing these essays about how lawyers should adapt, or maybe how journalists should adapt. But perhaps ironically, the question of how to adapt has been hitting our community first. Whether it's software engineering or AI research itself, AI capabilities in these areas are of course advancing very rapidly.

Our response to this moment matters beyond this community. The whole world is watching. If we simply roll over and accept that a lot of our work will be done by AI in the future, instead of setting clear boundaries, I think it will lead to an even stronger political backlash against AI than what we are seeing today. So I think this question is not just for us but for the whole world.

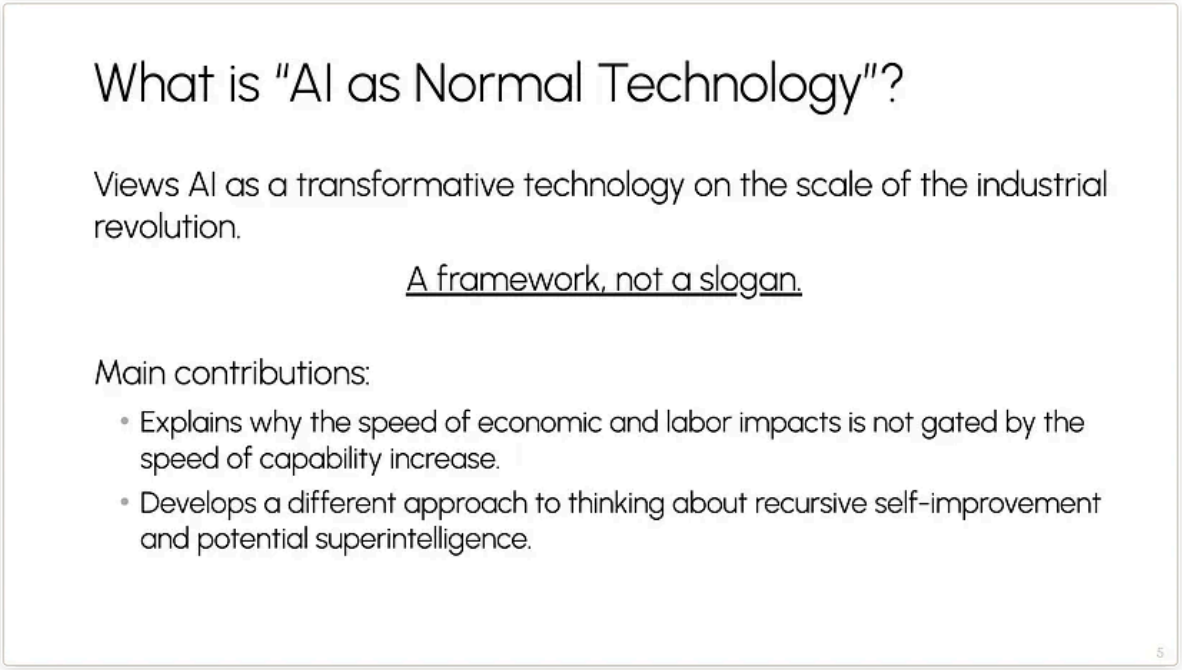

A battle of two narratives
If AI will replace our expertise:
Build wealth before AI obviates our skills.
If AI will amplify our potential:
Build skills, agency, taste, judgment.
The choice matters.
What if you choose the first path and the second narrative turns out to be correct?
4

From the beginning of AI, historically there have been these two battling narratives. In the past, the distinction was academic and philosophical, but now it has become an acutely practical question. Each one of us has to decide which camp we're in, or where on this spectrum we're in, because the practical consequences of believing in one versus the other are very, very different.

If you think this is a technology which in a few years is going to be able to replace everything we do today, then perhaps the correct response is to build wealth as quickly as possible before our skills become irrelevant. And this is the path that many have chosen in Silicon Valley. You may have heard of the "permanent underclass" meme.

On the other hand, if you believe, as I do, that this is a technology that will greatly amplify our potential, then now is the best time to build skills — especially the skills that are going to be complementary to what AI is doing and is going to be able to do — as well as to build all the things around it, such as agency and taste and judgment.

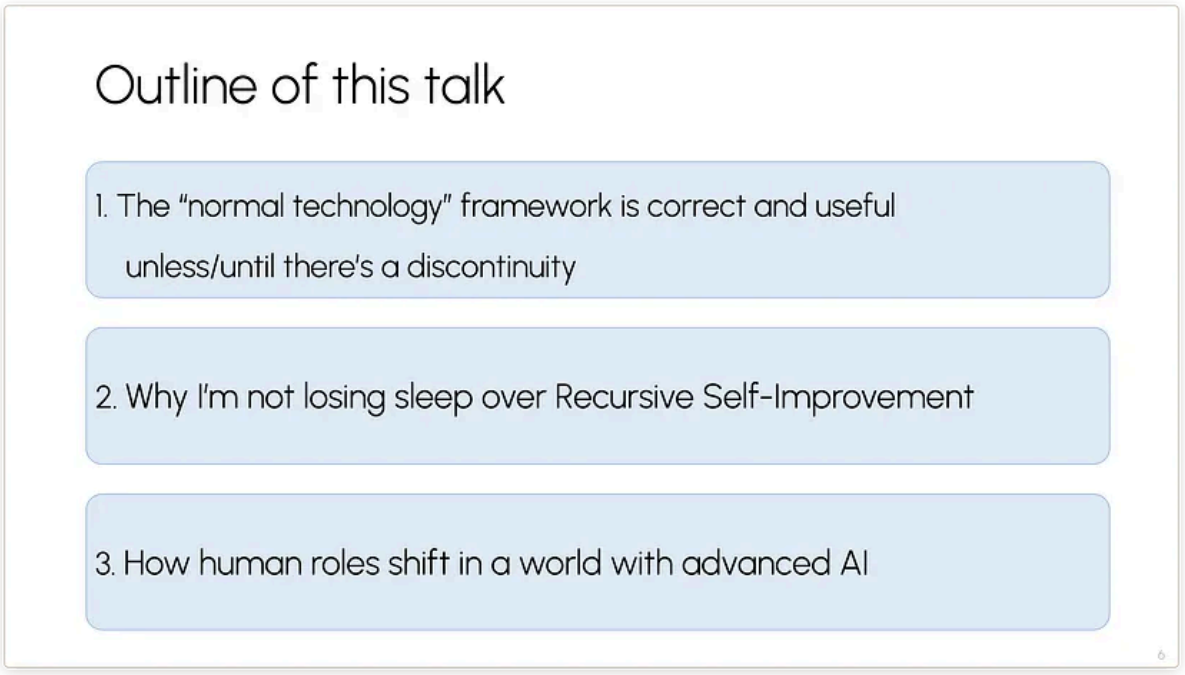
If you choose the first path, and it turns out that AI actually ends up being an amplifying technology as opposed to a replacing technology, then I would argue that over the next few years you've perhaps lost the best time in history to build these skills that will give us superpowers. That's why we all need to think about this question, even if we won't all land in the same place.


What is "AI as Normal Technology"?
Views AI as a transformative technology on the scale of the industrial revolution.
A framework, not a slogan.
Main contributions:
• Explains why the speed of economic and labor impacts is not gated by the speed of capability increase.
• Develops a different approach to thinking about recursive self-improvement and potential superintelligence.
5

AI as Normal Technology is the intellectual framework for my talk today. When we say AI is normal, we don't mean that it's just like a hammer or a toothbrush, some kind of mundane technology.

We acknowledge prominently in the essay that this is a transformative technology on the scale of the industrial revolution. We're not AI skeptics.

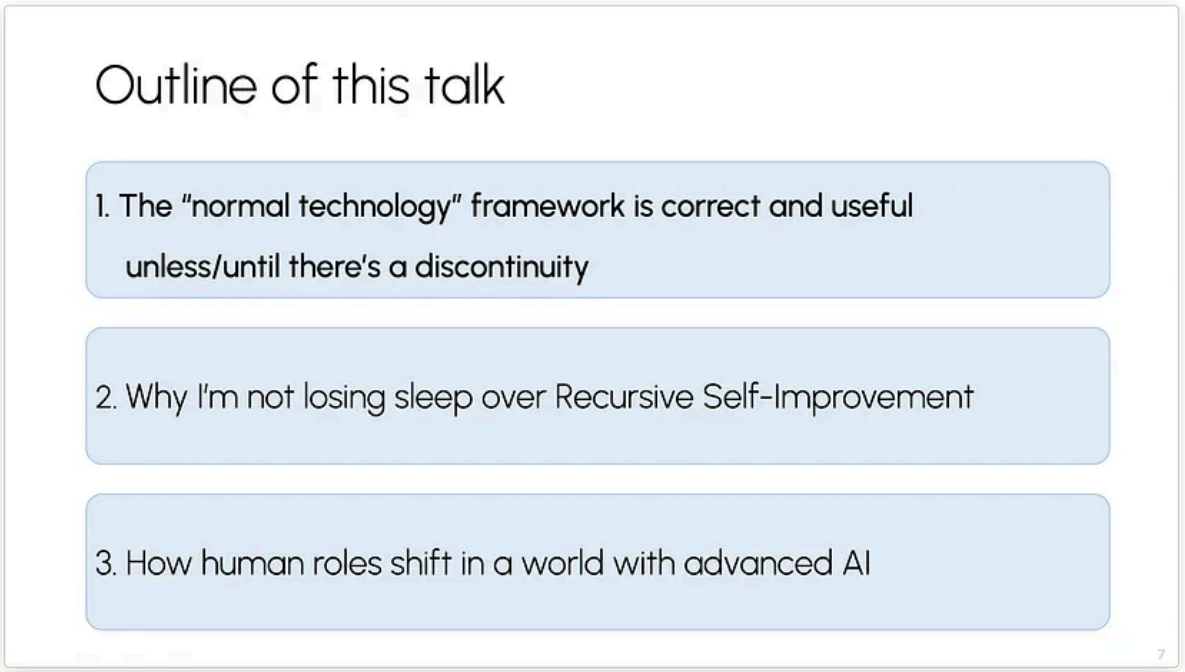
This is not a slogan. It's a framework — sort of a causal model how AI capabilities impact the economy and society. It's a 15,000-word essay and we're turning it into a book. And I mention that because people often hear the word normal and they assume they know what we mean, but that leads to misunderstandings.

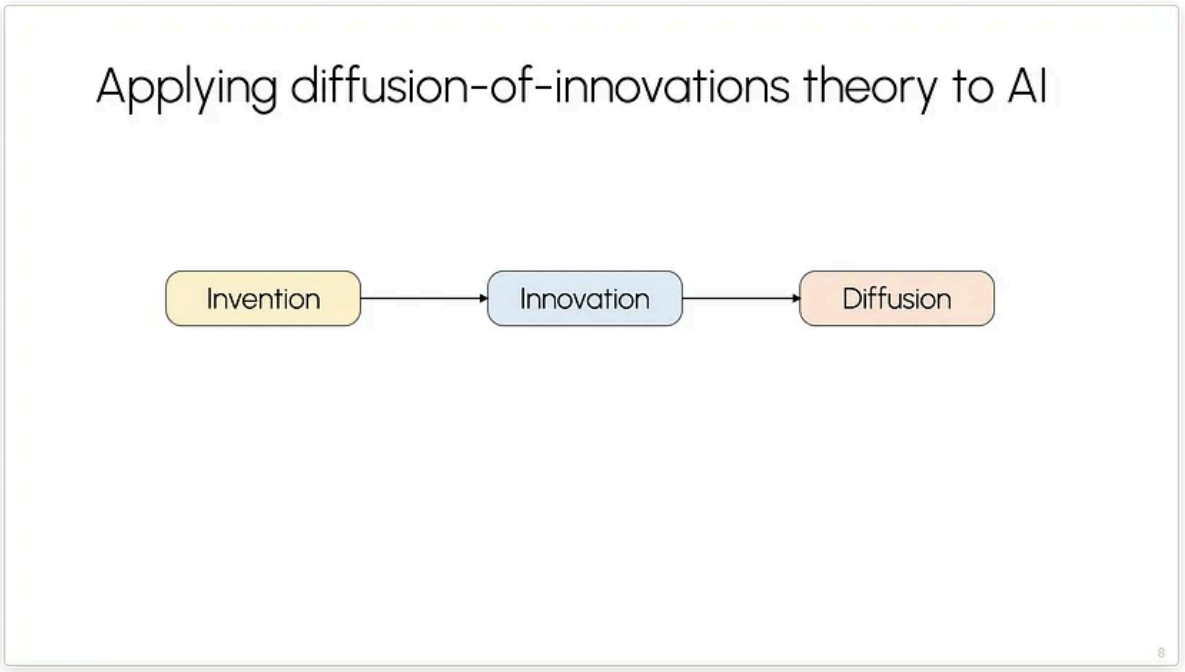

Outline of this talk
1. The "normal technology" framework is correct and useful unless/until there's a discontinuity
2. Why I'm not losing sleep over Recursive Self-Improvement
3. How human roles shift in a world with advanced AI
6

First, I'll argue that this framework is correct and useful as a way to think about AI's impacts, unless and until there is some future discontinuity — such as through recursive self-improvement — that leads to future impacts looking very different from past impacts.

Second, I will talk about why, even though we should take recursive self-improvement seriously, I'm not particularly losing sleep over it.

Third and finally, I want to be clear that I'm not saying that jobs of the future will be just like jobs of the present. A lot of adaptation will be needed. So I want to give some preliminary thinking on how I think our roles are going to change and how we can best adapt to the changes.


Outline of this talk
1. The "normal technology" framework is correct and useful unless/until there's a discontinuity
2. Why I'm not losing sleep over Recursive Self-Improvement
3. How human roles shift in a world with advanced AI
7



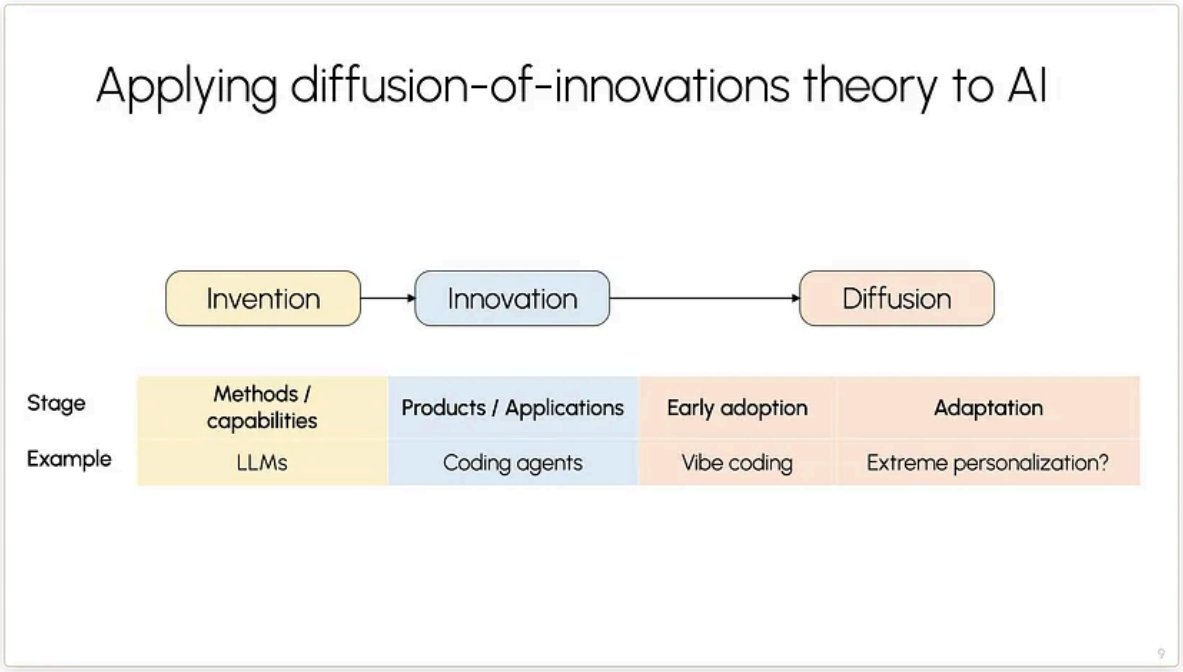
Powerful technologies of the past, like electricity, have been thoroughly studied, and we have good frameworks to understand how technological progress leads to economic impacts.

Invention: discovering the principles of electromagnetism, AC versus DC, etc.

Innovation: People don't use "electricity" directly. We use electrical appliances. Those had to be invented, so that is a kind of downstream innovation — that's the second phase of the framework.

Diffusion: This refers to the gradual process by which people start adopting innovations.

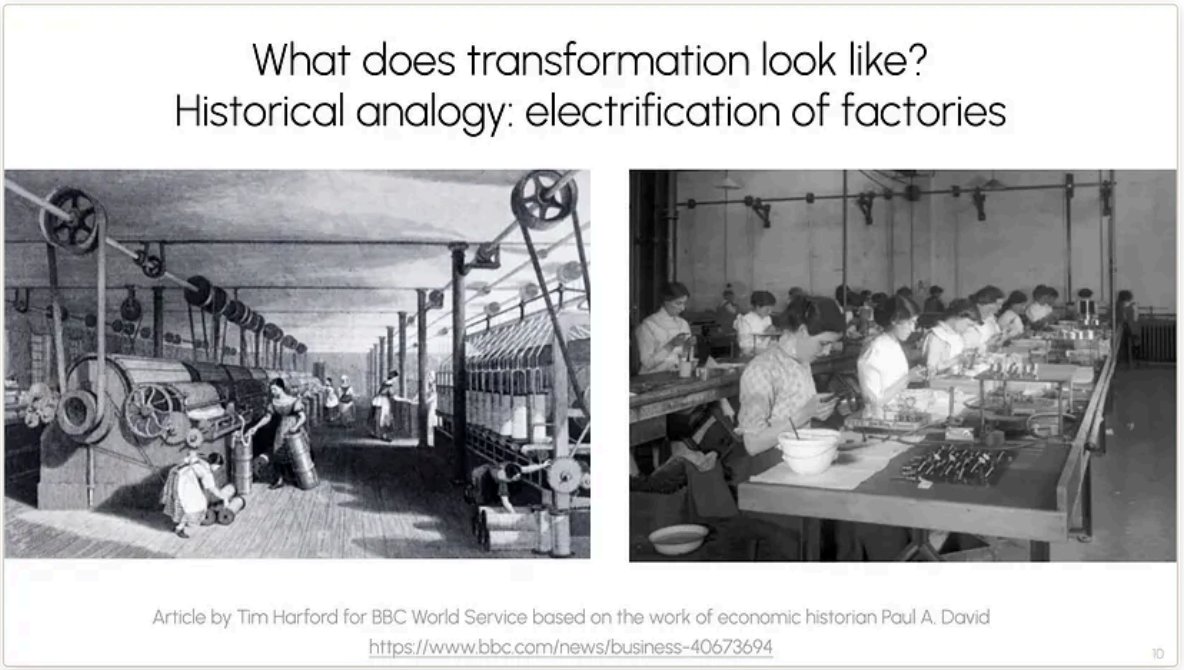
In our essay we apply this framework to AI and flesh it out into a four-part framework.



Here's the basic picture, illustrated with software engineering as an example.
Methods/capabilities: Models are rapidly improving.
Products/applications: We don't use LLMs directly. The reason they've been so influential in all of our work is because of coding agents. These are products that take those latent capabilities and turn them into something useful and usable for workers.
Early adoption: At first people were mostly trying vibe coding, and now we know that that's not really the best way to develop production software — so now we have more sophisticated ways of doing agentic engineering.
Adaptation: (or structural transformation) — the fourth and slowest phase. Much of my talk today is going to be about that. I claim that this stage takes decades. It has not really started yet, even in a field like software engineering, which is a relative early adopter of coding agents.

We don't know what the adaptation phase will look like — we can only speculate. Permit me to speculate for a minute. If it's going to be the case that coding agents are going to be able to create ten-million-line code bases in the future that are not full of bugs and security vulnerabilities, then it won't make a lot of sense for us to create one piece of software that billions of people should use. It'll make a lot more sense for software to be tailored to the needs of each individual or team. And that's what I mean by extreme personalization.

It’s not merely a technological change — that’s also a change for the industry. For instance: do we even need software companies anymore? Maybe software development will massively shift in-house, into the companies and teams that are actually using the software. Again, this is speculation, but the point is that it is this kind of organizational change, human change — that’s very slow, that takes decades — that will allow us to take advantage of the full potential of AI, whether it’s in software engineering or in any other field. So that’s one of the central insights of the essay. When we look at past technologies, this kind of change tends to be very slow.


What does transformation look like?
Historical analogy: electrification of factories
Article by Tim Harford for BBC World Service based on the work of economic historian Paul A. David
https://www.bbc.com/news/business-40673694
10

Before electricity, factories used to look like the picture on the left. A massive steam engine generated power and it was moved throughout the factory by mechanical gears and belts. So when electricity came along, factory owners tried to replace those steam boilers with electric generators. They thought it would be much more efficient. But this idea of a drop-in replacement did not work. We keep hearing that term in the context of AI agents today — that they will be drop-in replacements for human workers. That did not work in the case of electricity.

What actually worked, and what took 40 years to develop, is to recognize that electricity is a very different technology. It’s portable, so you can move the power to wherever you need it. That lets you reorganize the entire layout of the factory around the logic of the assembly line. And that required changing the way that workers are trained, hired, and fired, new labor laws, and so forth. So that’s the kind of organizational adaptation that it took in order to reap the benefits of electricity in factories.

Our claim is that this is the kind of process that we will go through for AI. A couple of decades from now, we will have fundamentally reorganized work. We don’t know

what that's going to look like, and that is the challenge in front of all of us. And that's not just a job for the AI companies to do, much like it wasn't the job of the electric utility to figure out how factories should be reorganized. In our view, this is the slowest of the four stages through which AI leads to economic impacts. Today, this process has not really gotten started.

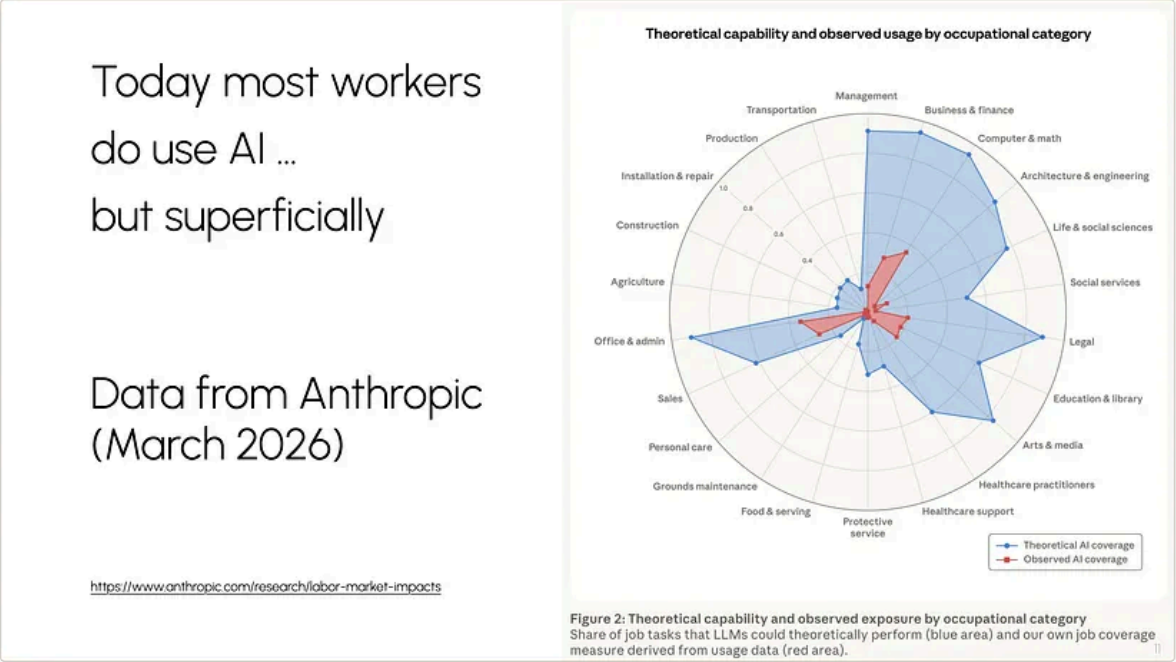
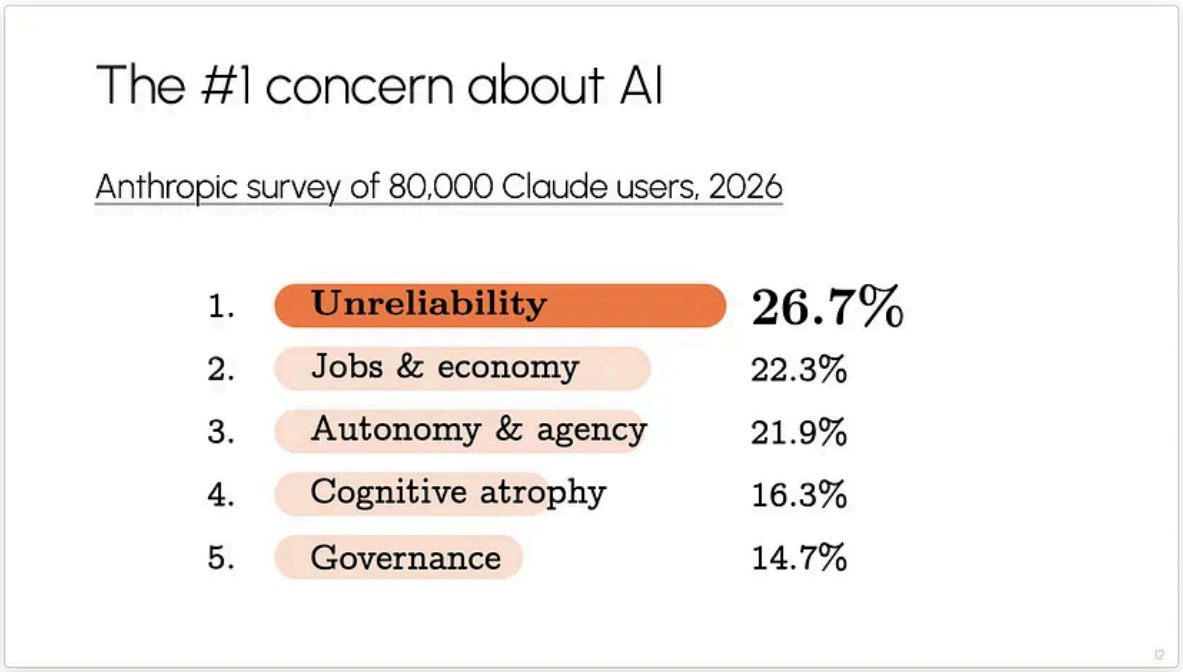


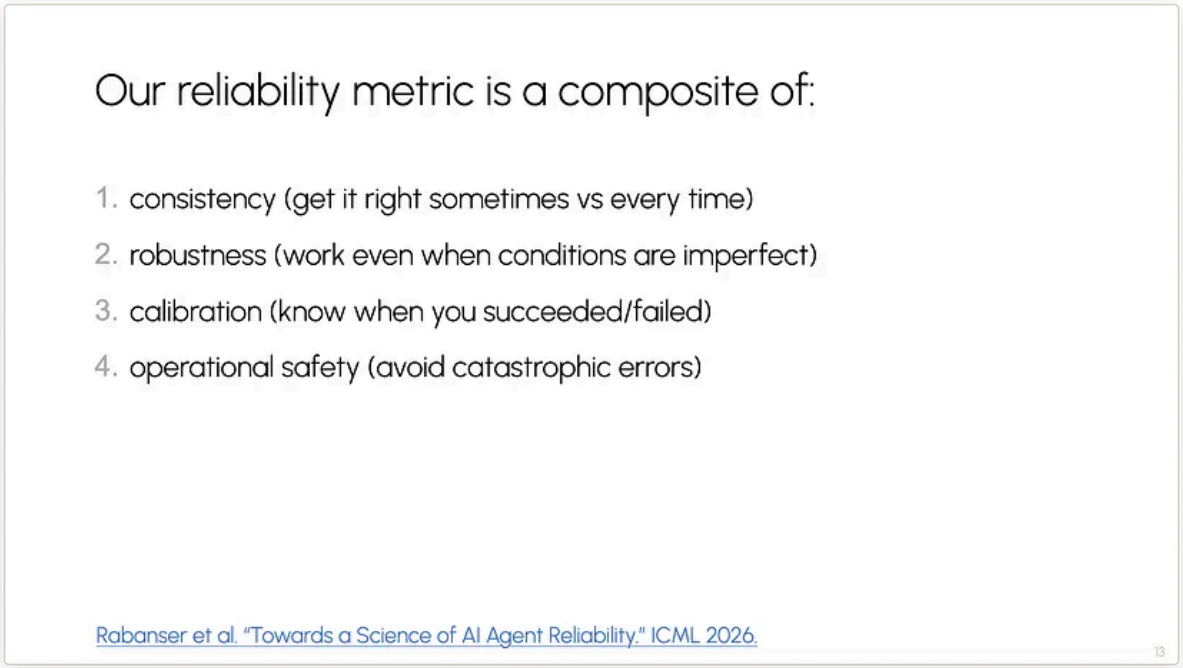
Figure 2: Theoretical capability and observed exposure by occupational category
Share of job tasks that LLMs could theoretically perform (blue area) and our own job coverage measure derived from usage data (red area).

Why is there a huge gap between what people in various occupations could be using AI for and what they're actually using it for? One reason could be that people are slow to adopt technology, and that's certainly part of our framework.

But we wondered if maybe the people who are deploying AI and are not having much success at it know something about the practical limitations of AI that the AI industry doesn't. Let's have a bit more humility about the relationship between capabilities and deployment.


The #1 concern about AI
Anthropic survey of 80,000 Claude users, 2026
1. Unreliability 26.7%
2. Jobs & economy 22.3%
3. Autonomy & agency 21.9%
4. Cognitive atrophy 16.3%
5. Governance 14.7%
12

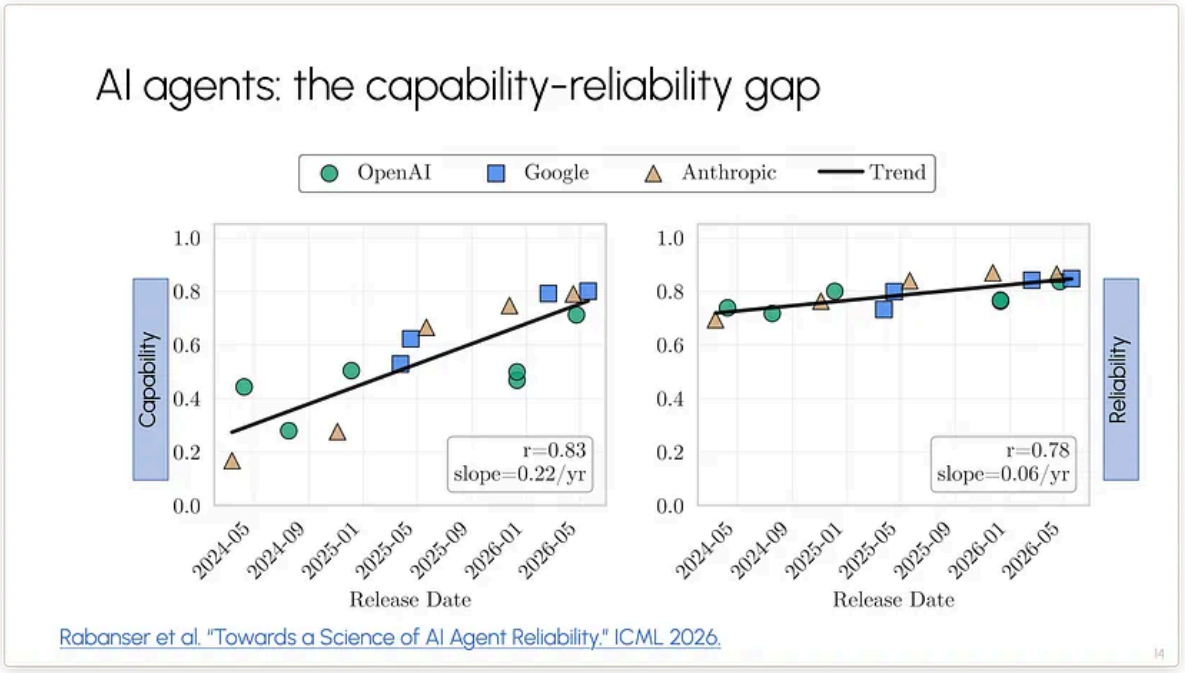
Given that the #1 concern people cite is reliability, we wanted to try to measure whether reliability, distinct from capability, is a barrier to the practical usefulness of AI agents.


Our reliability metric is a composite of:
1. consistency (get it right sometimes vs every time)
2. robustness (work even when conditions are imperfect)
3. calibration (know when you succeeded/failed)
4. operational safety (avoid catastrophic errors)
Rabanser et al. "Towards a Science of AI Agent Reliability." ICML 2026.
13

We looked at 10-12 reliability metrics and clustered them into four dimensions. **Consistency:** Suppose we hear that an AI agent has a 70% accuracy. Does this mean it works on 70% of the tasks, but on the ones that it does, it does so every time? That's great for deployment — you can deploy it on that subset of the tasks. Or does it mean that on any given task it might unpredictably fail with a 30% probability? That's pretty useless from a deployment perspective. Perhaps shockingly, none of the agent benchmarks that we looked into make a distinction between these two. Both of

these are represented as 70% accuracy.
Robustness: We looked at robustness: what happens when the environment changes a little bit?
Calibration: Can the agent look back at its transcript and tell if it performed the task correctly?
Operational safety: When it does fail, is it recoverable, or is it something like deleting the production database?

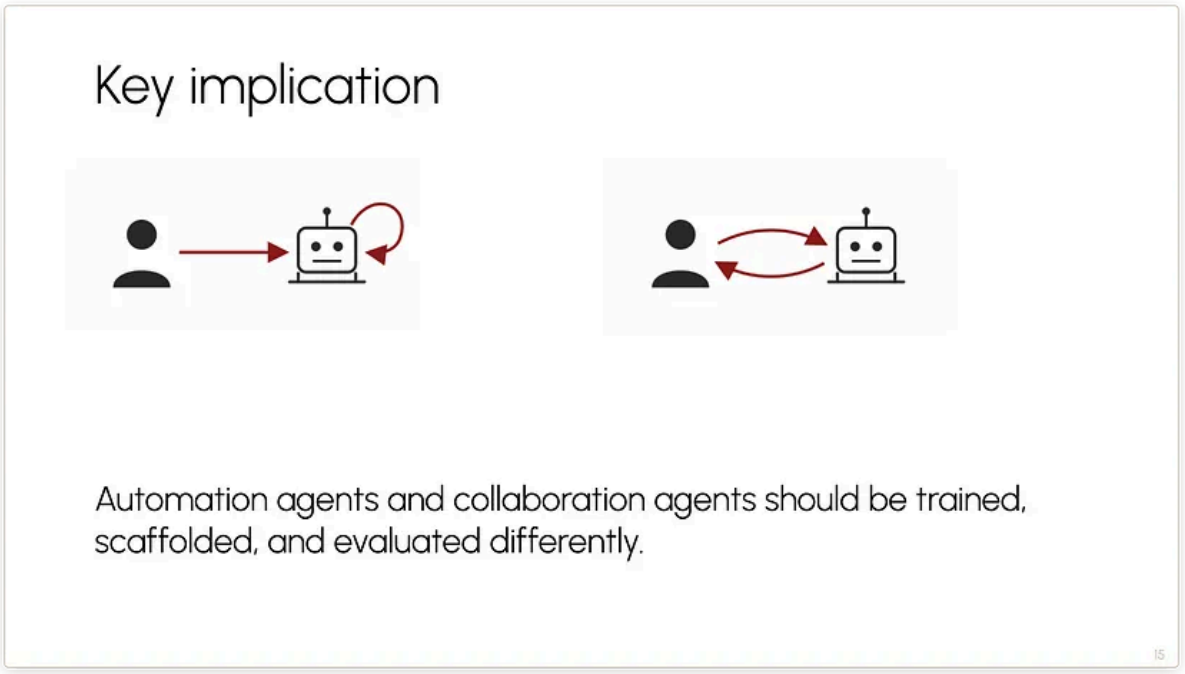
For a human worker, if we think of someone as being competent at a job, it's all of these things, not just accuracy. But it turns out we were measuring agents only on accuracy.



We measured capability and reliability using two complementary benchmarks, for models from these 3 frontier AI companies that were released over the last 24 months or so.

This is a period during which accuracy or capability shot up dramatically (left).

But reliability (right) only increased by five or ten percentage points.


Key implication
Automation agents and collaboration agents should be trained, scaffolded, and evaluated differently.
15

There are a bunch of implications but let me highlight one. Right now the industry is not treating automation and collaboration agents differently. If you run an agent in headless mode, I guess that becomes an automation agent. That is not a good way to look at it, because properties like reliability that are very important for automation agents can actually be a hindrance for a collaboration agent that you might be using to improve your creative writing or something like that. For that kind of agent, you *don't* want it to behave like a robot that does the same thing every time. You want it to be creative and explore different possibilities and be unpredictable.

Scaffolds and even the post-training of models should be different based on whether it is supposed to be driving a collaboration agent or an automation agent.

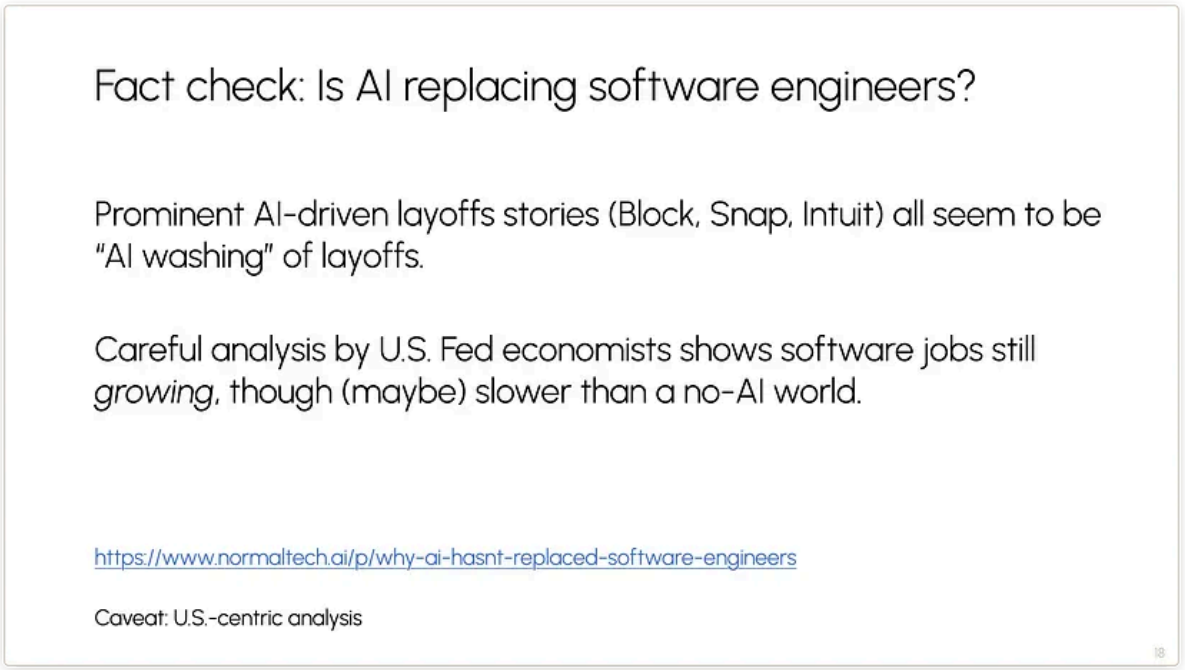
Prediction: collaboration agents will continue to be more successful than automation agents.
Many hasty deployments are being rolled back.
High-stakes
General purpose
Automated
The AI agent trilemma:
You can only have two

My hope is that reliability will continue to improve and automation will become easier over time. But for now I think collaboration agents will continue to be much more successful. Many companies that hastily rushed to automate business processes using agents are recognizing the limits and costs, and even legal liabilities — like when an agent deletes production data.

For the time being, you can have only two out of these three properties in agents: **general-purpose** (a language model based agent that can be instructed to do different tasks rather than purpose-built for a task like traditional software); deployed in **high-stakes** scenarios, and **automated**.

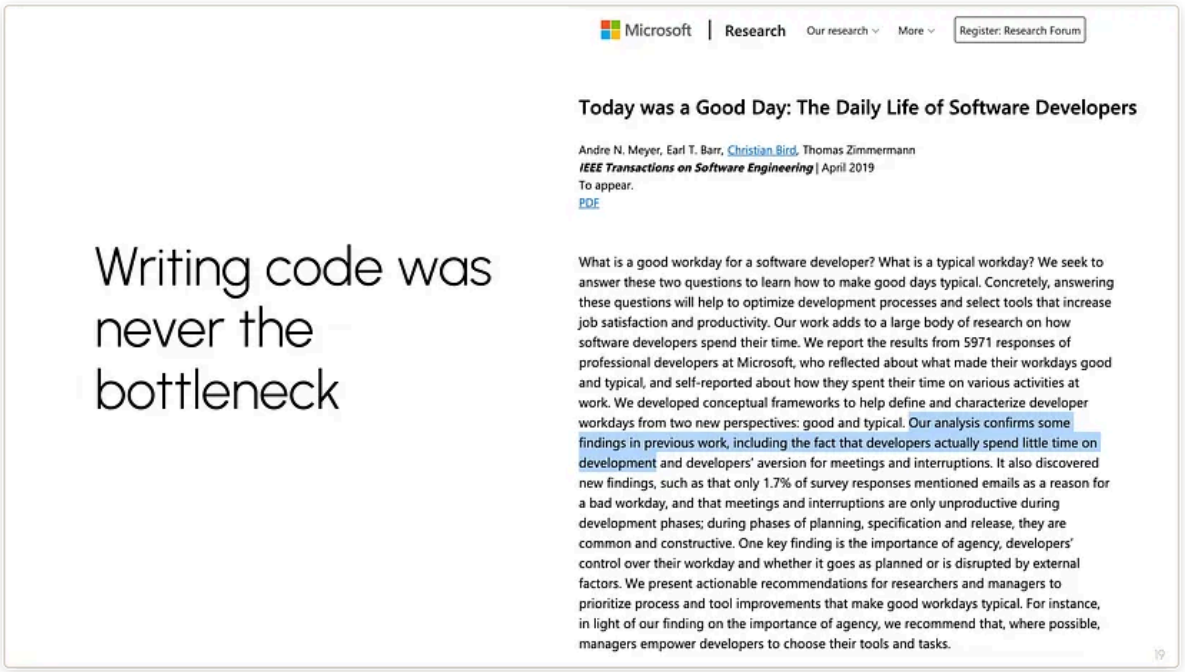
This is one of the reasons that I think that for now, AI remains much more of a collaboration technology than a technology that automates workers away.

Let’s take software engineering as a case study. It’s a good leading indicator because coding agents have been particularly rapidly adopted.

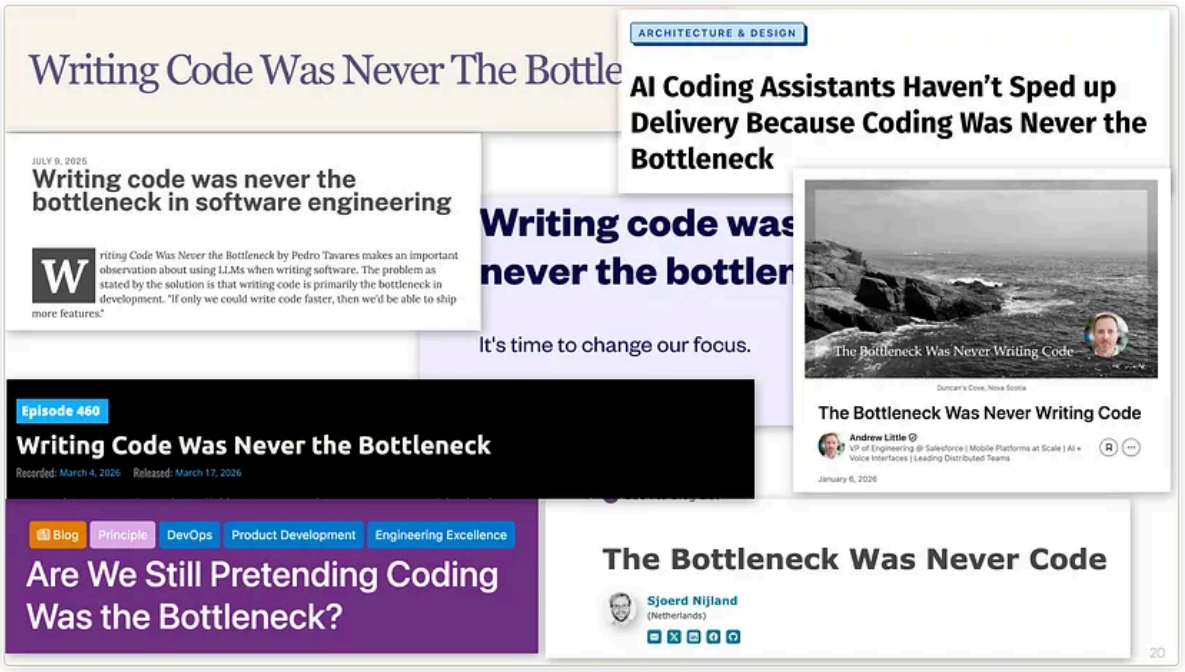

Fact check: Is AI replacing software engineers?
Prominent AI-driven layoffs stories (Block, Snap, Intuit) all seem to be "AI washing" of layoffs.
Careful analysis by U.S. Fed economists shows software jobs still growing, though (maybe) slower than a no-AI world.
https://www.normaltech.ai/p/why-ai-hasnt-replaced-software-engineers
Caveat: U.S.-centric analysis
18

You might think: okay, we can’t completely automate away software engineering. But if agents make software engineers ten times more productive, then we need ten times fewer software engineers. Isn’t that an obvious consequence?

Well, that is completely contradicted by the data. We looked at this in a follow-up essay. In every case we looked at, the company was under financial pressure, and it turns out to be more convenient to blame AI for the layoffs instead.

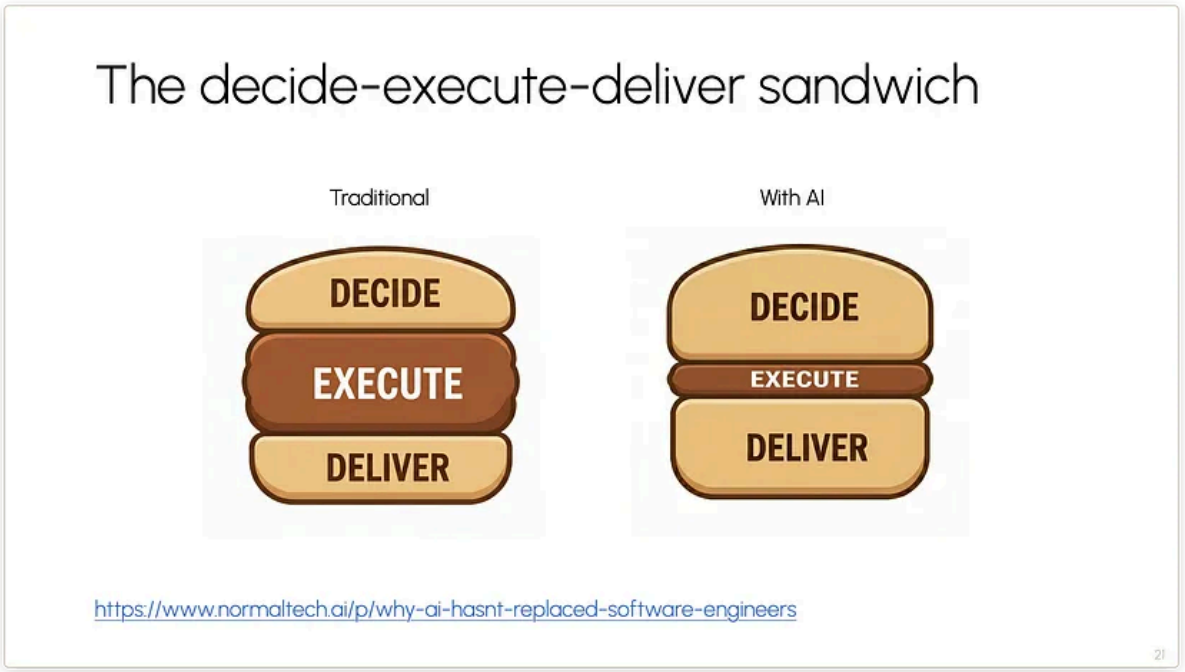

Writing code was never the bottleneck
Microsoft | Research
Our research
More
Register: Research Forum
Today was a Good Day: The Daily Life of Software Developers
Andre N. Meyer, Earl T. Barr, Christian Bird, Thomas Zimmermann
IEEE Transactions on Software Engineering | April 2019
To appear.
PDF
What is a good workday for a software developer? What is a typical workday? We seek to answer these two questions to learn how to make good days typical. Concretely, answering these questions will help to optimize development processes and select tools that increase job satisfaction and productivity. Our work adds to a large body of research on how software developers spend their time. We report the results from 5971 responses of professional developers at Microsoft, who reflected about what made their workdays good and typical, and self-reported about how they spent their time on various activities at work. We developed conceptual frameworks to help define and characterize developer workdays from two new perspectives: good and typical. Our analysis confirms some findings in previous work, including the fact that developers actually spend little time on development and developers' aversion for meetings and interruptions. It also discovered new findings, such as that only 1.7% of survey responses mentioned emails as a reason for a bad workday, and that meetings and interruptions are only unproductive during development phases; during phases of planning, specification and release, they are common and constructive. One key finding is the importance of agency, developers' control over their workday and whether it goes as planned or is disrupted by external factors. We present actionable recommendations for researchers and managers to prioritize process and tool improvements that make good workdays typical. For instance, in light of our finding on the importance of agency, we recommend that, where possible, managers empower developers to choose their tools and tasks.
19

Why is AI not replacing software engineers so far? We've known for a while — this is a paper from 2019 — that writing code is not really the bottleneck.


Writing Code Was Never The Bottle
ARCHITECTURE & DESIGN
AI Coding Assistants Haven't Sped up Delivery Because Coding Was Never the Bottleneck
JULY 9, 2025
Writing code was never the bottleneck in software engineering
Writing Code Was Never the Bottleneck by Pedro Tavares makes an important observation about using LLMs when writing software. The problem as stated by the solution is that writing code is primarily the bottleneck in development. "If only we could write code faster, then we'd be able to ship more features."
Writing code was never the bottler
It's time to change our focus.
The Bottleneck Was Never Writing Code
Andrew Little
Episode 460
Writing Code Was Never the Bottleneck
Recorded: March 4, 2026 Released: March 17, 2026
Blog
Principle
DevOps
Product Development
Engineering Excellence
Are We Still Pretending Coding Was the Bottleneck?
The Bottleneck Was Never Code
Sjoerd Nijland
(Netherlands)
20

Over the last year, as software engineers started adopting coding agents and started to recognize that it doesn't seem to be cutting down on the amount of their work, there have been many blog posts rediscovering the fact that writing code is not a bottleneck. Here is a small sample.

So what actually is the bottleneck?



This framework is our answer to the question.

The decide layer: understanding customer requirements, developing the specification, planning, etc. That is not getting compressed by AI.

The execute layer: the actual coding and debugging. This is getting compressed, but it was only maybe one-third of the work to begin with.

The deliver layer: Understanding your code deeply enough to be accountable for what you release; carrying out integration into customer systems, maintenance, testing, etc. This layer is not getting compressed either.

In fact, the first and third layers are arguably expanding as AI compresses the middle layer — and I'll come back to that point.

A useful metaphor: A crane operator

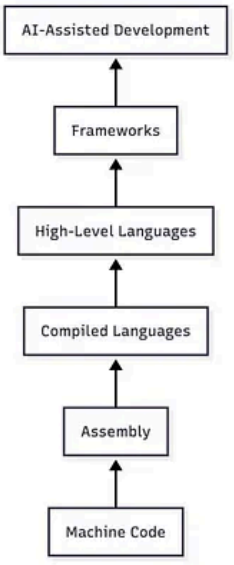
AI amplifies human potential but the human remains in control.

22

I think it is already the case in software engineering, and will increasingly be the case in many professions, that we can think of knowledge workers as similar to a crane operator or a forklift operator. The machine greatly amplifies the human potential to do physical work — it's doing all the heavy lifting, but the person still remains in control. And I think this is what is happening with cognitive work. Machines are going to increasingly do the cognitive heavy lifting, but the person still remains in control.

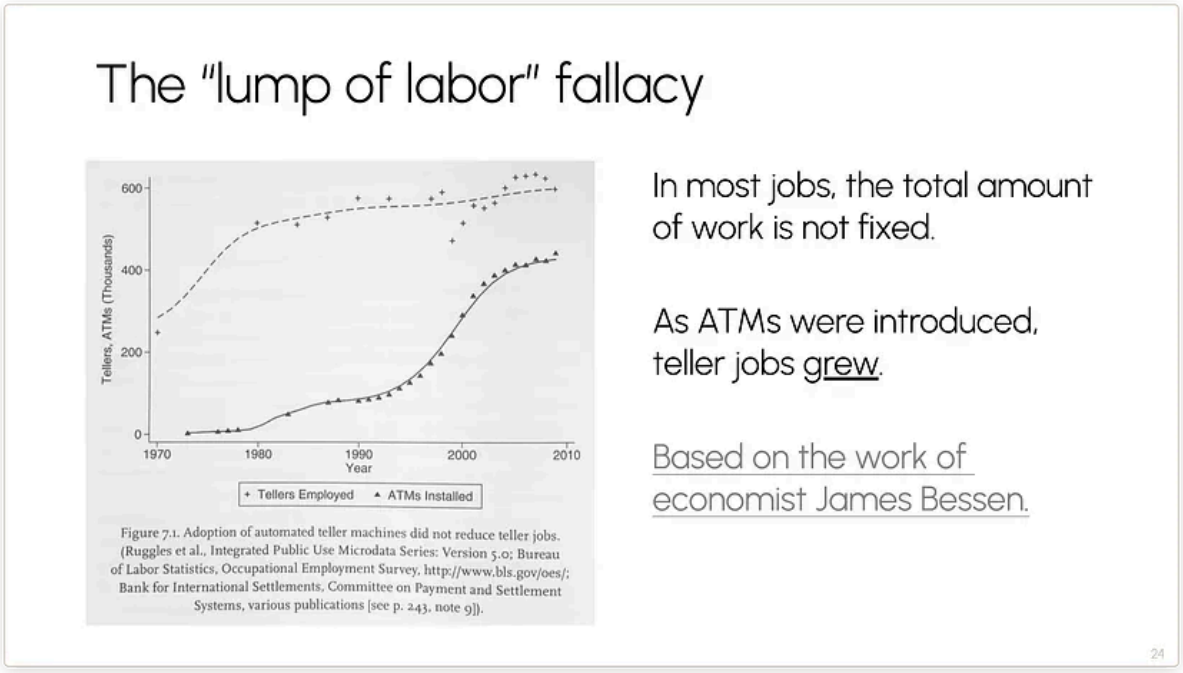
The entire job gets reconceptualized as being about operating the machine, understanding the machine, and controlling the machine, as opposed to doing the cognitive work ourselves.

Change is constant in software engineering



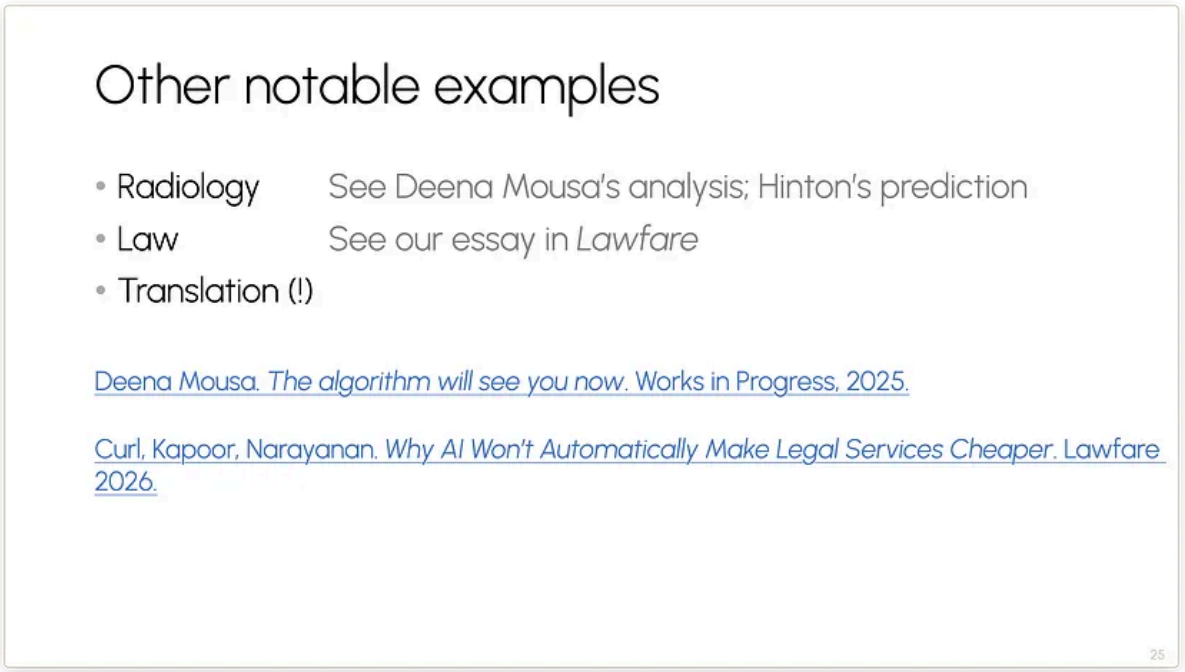
23

All this might seem like a dramatic change, but in a sense it is only a continuation of what has been happening over and over and over in software engineering. Starting from the days of machine code, we've had many waves of technology, each of which gives us nearly an order of magnitude increase in productivity. And far from decreasing the demand for software engineers, during the time that we've climbed up this ladder, the amount of software engineering employment has increased by a factor of something like 10,000. And that's simply because the amount of code that there is to write has gone up by orders and orders of magnitude.


The "lump of labor" fallacy
600
400
200
0
Tellers, ATMs (Thousands)
1970
1980
1990
2000
2010
Year
+ Tellers Employed
▲ ATMs Installed
Figure 7.1. Adoption of automated teller machines did not reduce teller jobs. (Ruggles et al., Integrated Public Use Microdata Series: Version 5.0; Bureau of Labor Statistics, Occupational Employment Survey, http://www.bls.gov/oes/; Bank for International Settlements, Committee on Payment and Settlement Systems, various publications [see p. 243, note 9]).
In most jobs, the total amount of work is not fixed.
As ATMs were introduced, teller jobs grew.
Based on the work of economist James Bessen.
24

Economists have found this repeatedly. You might have heard the term Jevons' paradox; I like the term "lump-of-labor fallacy".

ATMs made it economically feasible for banks to open lots of regional branches. And those branches still needed human tellers to handle the things ATMs couldn't. So, paradoxically, employment actually grew.


Other notable examples
• Radiology See Deena Mousa's analysis; Hinton's prediction
• Law See our essay in Lawfare
• Translation (!)
Deena Mousa. The algorithm will see you now. Works in Progress, 2025.
Curl, Kapoor, Narayanan. Why AI Won't Automatically Make Legal Services Cheaper. Lawfare 2026.
25

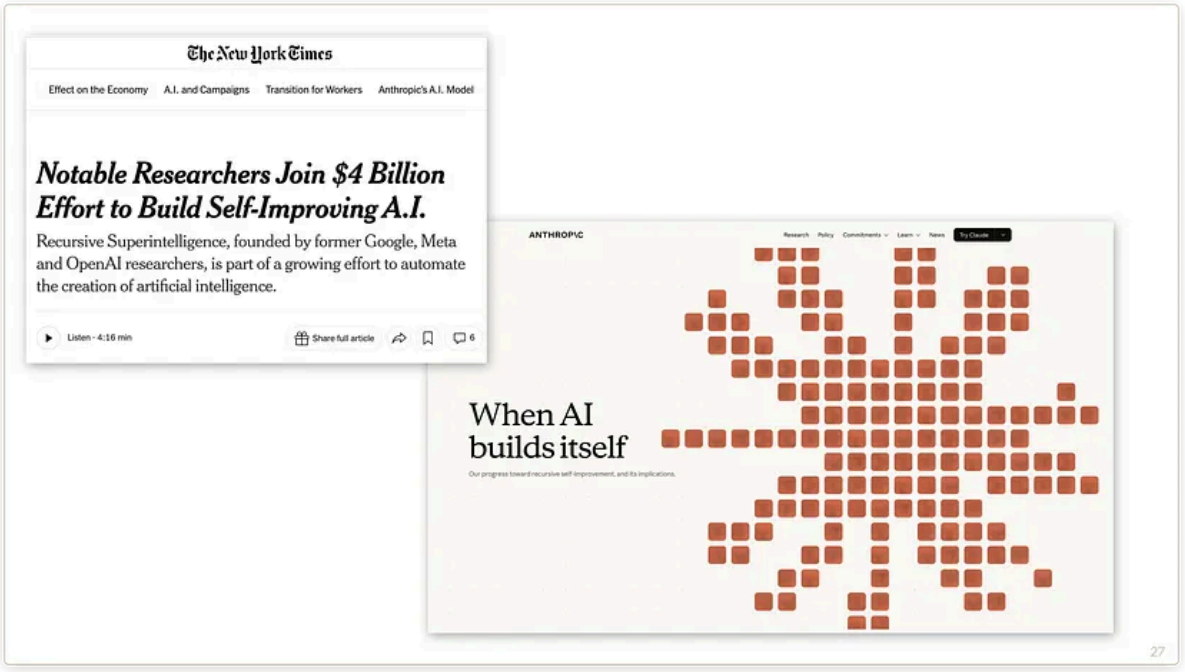
Geoff Hinton made the famous prediction a decade ago that radiologists would be basically extinct in five years. But it turns out radiology employment has in fact grown. And it's not because radiologists are rejecting AI. They are actually enthusiastically adopting it. One reason for job growth is that when a task gets faster and cheaper to perform, there's more demand for it.

We've written a paper looking at how lawyers should adapt. There's a lot in there, but one simple point is that AI has made it a lot easier to file lawsuits. And this means more work for lawyers. We might be unhappy about this if we don't like living in a litigious society, but from the point of view of employment for lawyers, this is great news.

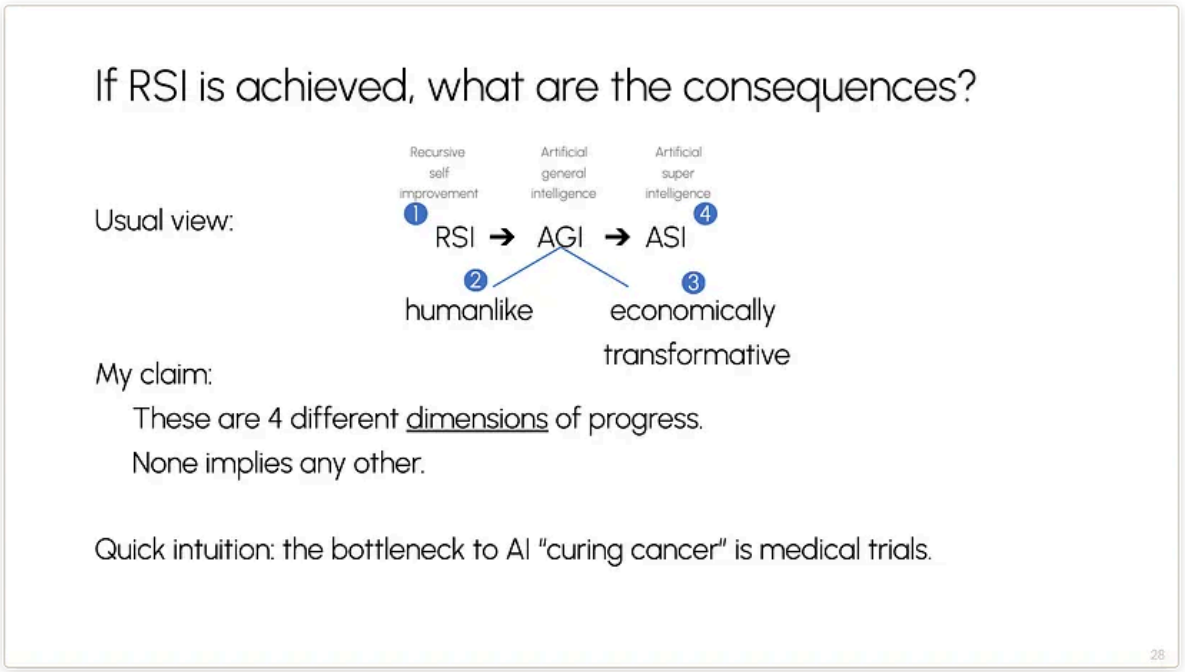
Translation is a more extreme example which kind of blows my mind. AI was almost at human parity nearly a decade ago. Yet the employment of human translators has remained more or less stable, and is projected to remain stable over the next decade. There are many reasons for that, but one reason is that there's really no ceiling to the amount of things you can translate and the number of different languages you can translate them into.

Outline of this talk
1. The "normal technology" framework is correct and useful unless/until there's a discontinuity
2. Why I'm not losing sleep over Recursive Self-Improvement
3. How human roles shift in a world with advanced AI
26

I've argued that so far, our framework is very consistent with the evidence. Now let's talk about the possibility that everything I've said so far will be obviated because some kind of takeoff or singularity will be achieved, and at *that* point there will be nothing left for us to do.


The New York Times
Effect on the Economy
A.I. and Campaigns
Transition for Workers
Anthropic's A.I. Model
Notable Researchers Join $4 Billion Effort to Build Self-Improving A.I.
Recursive Superintelligence, founded by former Google, Meta and OpenAI researchers, is part of a growing effort to automate the creation of artificial intelligence.
Listen · 4:16 min
Share full article
ANTHROPIC
When AI builds itself
27

Many companies have said, including these two, that they are racing towards recursive self-improvement. And they are serious companies — I take them seriously.



Suppose, some time in the next year, recursive self-improvement is achieved. What are the consequences? The slide shows the usual view in the AI community. Note that AGI a famously slippery term. There are two main groups of definitions. One is that AGI will be humanlike in a range of cognitive dimensions, and the other is that it will be capable of performing a range of economically valuable tasks.

My claim is that this view — that treats AGI and ASI as near-automatic consequences of RSI — is a little bit silly. These are four different dimensions of progress. None of these dimensions implies any of the others.

I'll talk about the details on the next few slides, but one small bit of intuition for now. One of the things that's associated with superintelligence is that it will cure cancer and other diseas. But we know that the hard part of developing medical treatments is clinical trials that often require thousands of people and 10-15 years. That's an example of the fact that the bottlenecks to superintelligence are external. It is not something that you can solve in the lab through purely computational processes. So the idea that recursive self-improvement will automatically lead to superintelligence, like many commentators assume — clearly there is something missing in terms of the causal chain.

AGI discourse suffers from the Hawaii problem

A term that once described a cluster of distant goals becomes unhelpful as we make progress toward them.

See Helen Toner's essay:
https://helentoner.substack.com/p/the-term-agi-is-almost-useless-at

29

Early on in the history of AI, these were all distant goals, so it was okay that we didn't clearly distinguish between different dimensions of progress. But now it's becoming a real problem — it's leading to confused discourse.

As an analogy, suppose we are early explorers and we hope to go to Hawaii one day. It's okay that we have one single term for this group of islands. But as our ship gets closer, we'd better be able to talk about these islands with different words. Otherwise, we're going to confuse ourselves about where we are and where we're going.

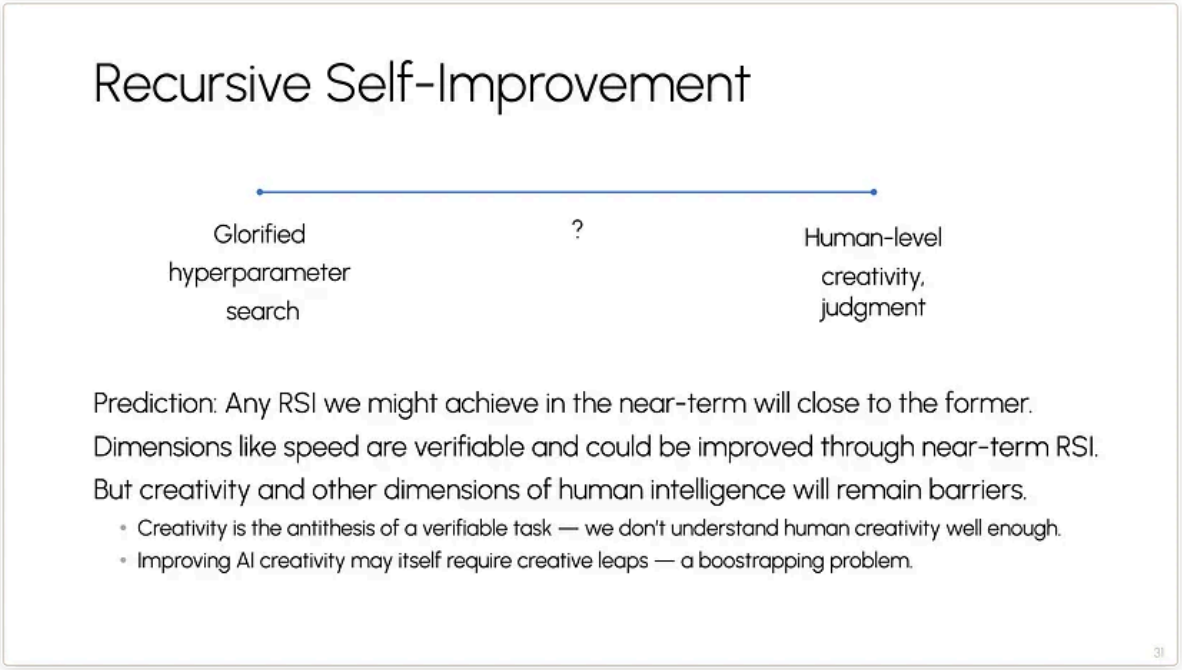
Helen Toner, among others, has made similar points.

Four dimensions of advanced AI

Recursive Self-Improvement	Humanlike AI
Economically transformative AI	Super-intelligence

30

Now let's discuss each of the four dimensions in detail, starting with Recursive Self-Improvement.


Recursive Self-Improvement
Glorified hyperparameter search
?
Human-level creativity, judgment
Prediction: Any RSI we might achieve in the near-term will close to the former.
Dimensions like speed are verifiable and could be improved through near-term RSI.
But creativity and other dimensions of human intelligence will remain barriers.
• Creativity is the antithesis of a verifiable task — we don't understand human creativity well enough.
• Improving AI creativity may itself require creative leaps — a boostrapping problem.

Suppose a company claims they have built RSI — they have built an AI system which built its own successor. What does that actually mean?

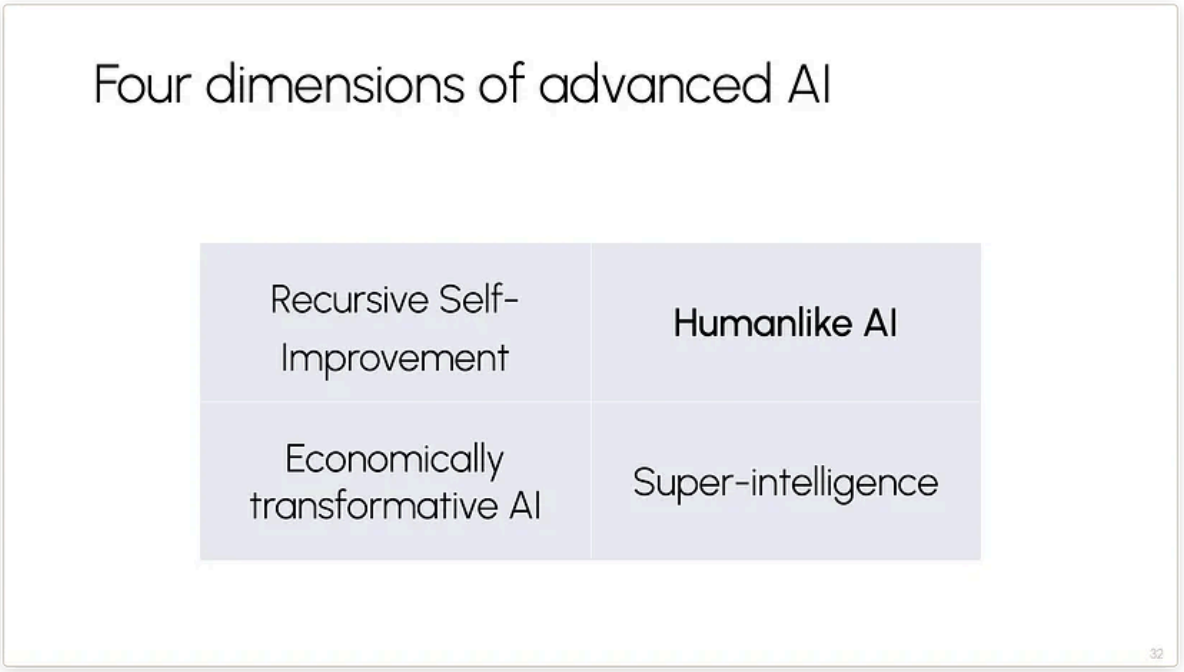
On the one hand, maybe it means an LLM spit out a whole bunch of ideas for how to tweak the architecture or the data pipeline or whatever else, and they automatically tested it and kept the improvements that worked. That's just glorified hyperparameter search. We've had systems like AutoML for a long time — I looked it up, and even Schmidhuber has published about this a long time ago. We're not calling that superintelligence. So that's one end of the spectrum.

It's very different from the other extreme, where you can imagine the company actually manages to replace the creativity and intelligence of the human AI researchers — not just one researcher, and not just researchers working at the company, but the entire worldwide community of hundreds of thousands of people whose innovations are all going into improving AI systems. So when people are talking about RSI, it's not clear which of these they mean.

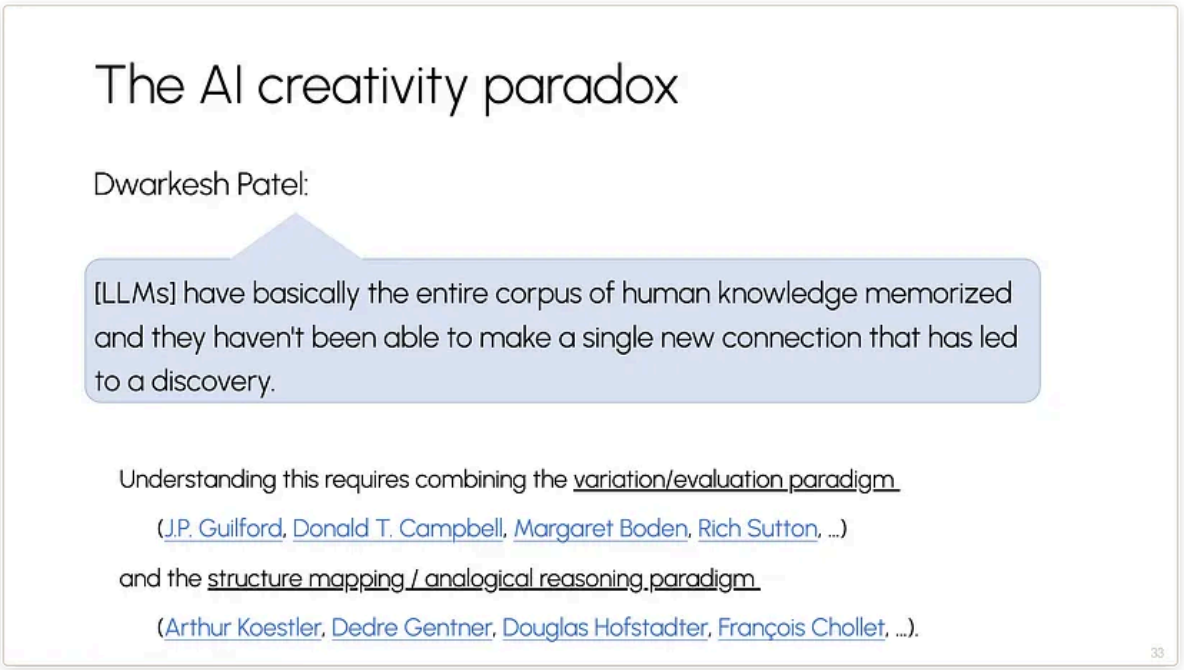
The reason for my prediction is simple: AI is still in a state where it is much better at verifiable tasks than non-verifiable tasks. Some dimensions of AI performance, like speed and efficiency, are verifiable, and it is possible that they could be greatly improved through near-term RSI.

However, creativity, among other dimensions of intelligence, is the epitome of an unverifiable task.

We don’t even know clearly how to test AI creativity. We don’t understand human creativity well enough from a cognitive science and neuroscience perspective to be able to have clear tests for it.



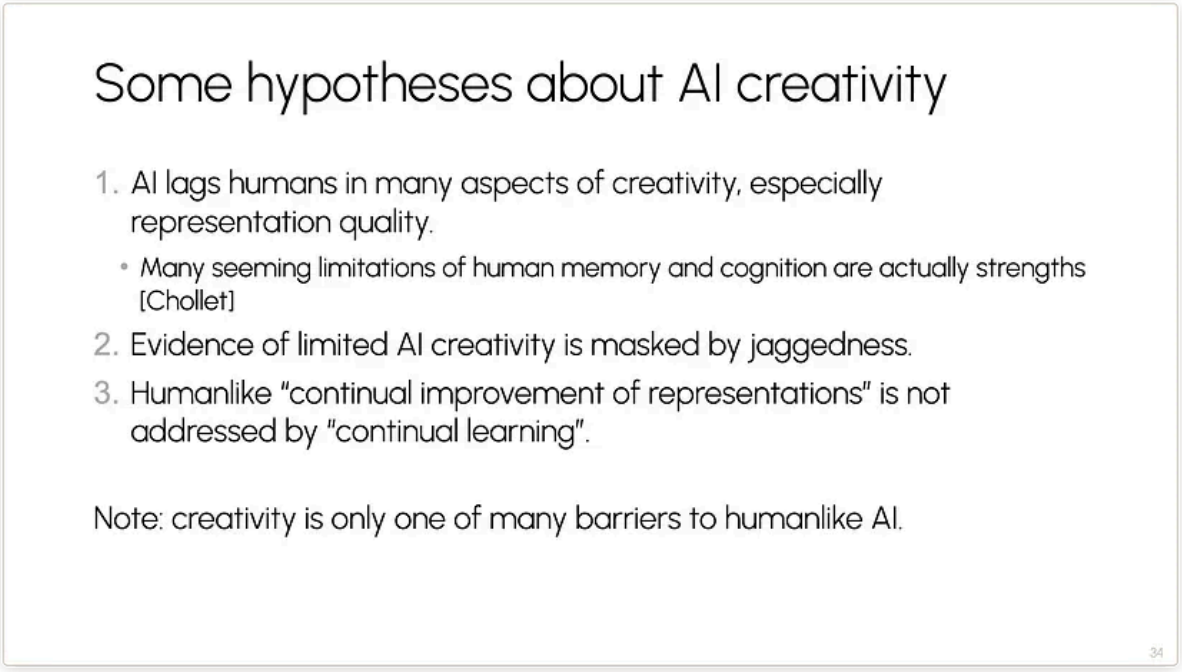
Let’s go a bit deeper into AI creativity as an illustration of the barriers to humanlike AI.


The AI creativity paradox
Dwarkesh Patel:
[LLMs] have basically the entire corpus of human knowledge memorized and they haven't been able to make a single new connection that has led to a discovery.
Understanding this requires combining the variation/evaluation paradigm
(J.P. Guilford, Donald T. Campbell, Margaret Boden, Rich Sutton, ...)
and the structure mapping / analogical reasoning paradigm
(Arthur Koestler, Dedre Gentner, Douglas Hofstadter, François Chollet, ...).
33

This quip from podcaster Dwarkesh Patel is a good illustration of how AI creativity lags human creativity. There might be a few counterexamples, especially in narrow domains like Erdos problems. But think about famous “Eureka moments” — finding surprising connections between seemingly unrelated fields or problems — that we

associate with great feats of human scientific invention and creativity. LLMs seem to be nowhere close to that.

To understand why, I spent a good amount of time immersed in the cognitive science literature. I won't go into the details, but seems like there are two literatures looking at this in slightly different ways that haven't really been talking to each other, for reasons I don't fully understand. Based on what I've been reading, I want to present some hypotheses about AI creativity.


Some hypotheses about AI creativity
1. AI lags humans in many aspects of creativity, especially representation quality.
• Many seeming limitations of human memory and cognition are actually strengths [Chollet]
2. Evidence of limited AI creativity is masked by jaggedness.
3. Humanlike "continual improvement of representations" is not addressed by "continual learning".
Note: creativity is only one of many barriers to humanlike AI.
34

Representation quality is absolutely fundamental to cognition. Think about deep learning — what a big leap that was in the quality of representations behind perception. So I do think AI has more or less caught up in representation quality when it comes to *perception*, but it has not caught up when it comes to representations that we use as humans for *creativity and reasoning*.

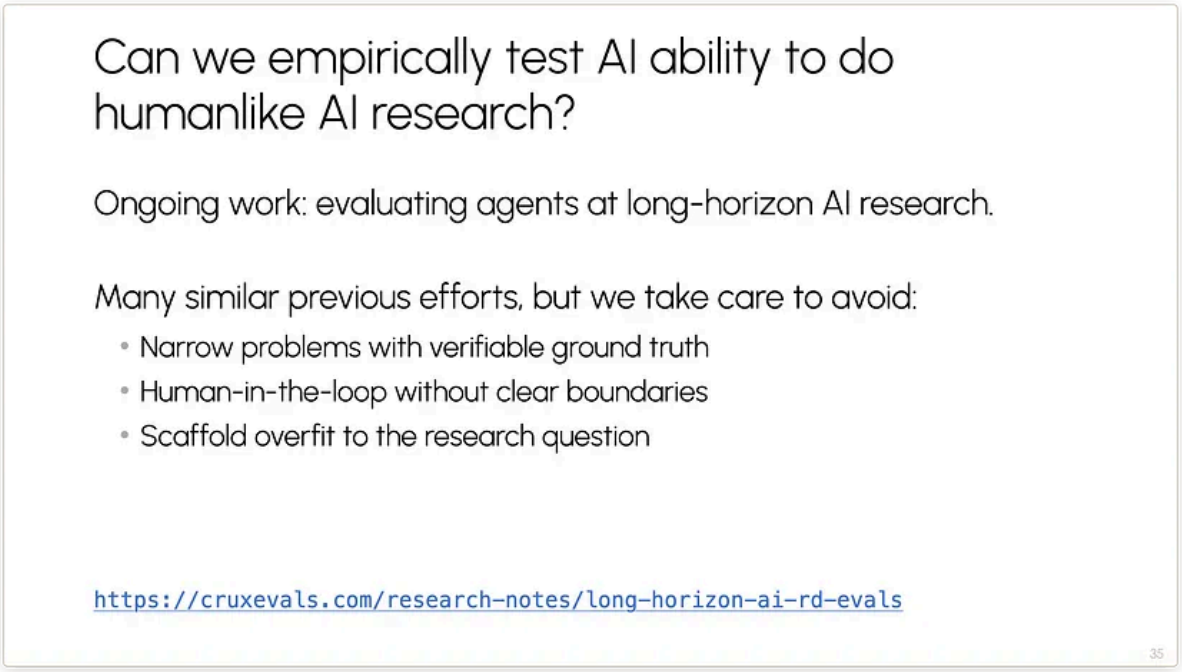
François Chollet in particular has talked about the fact that human representations that underpin our creativity seem to exhibit a kind of extreme compositionality, where everything is built up from a few "atoms of meaning". Many seeming limitations of human working memory and information processing actually turn out to be strengths, because they force us to come up with these extremely efficient representations.

Because LLMs are much better than humans in certain dimensions — like memorizing and retrieving stored patterns that might enable creativity — we're not able to elicit these important real limitations in studies that try to compare human and LLM creativity.

Furthermore, we humans can do something beautiful when we're creatively thinking about a problem: we improve our representations related to the problem in real time — at inference time, if you will. That enables us to sleep on it, improve our representations, and come back the next day much more efficient at solving that particular problem. I'm sure we've all experienced this. But it is something that today's AI systems are not able to do.

I assumed naively that the continual learning people would be all over this, but when I started looking into that literature, it appears to me that that's not the case. Continual learning is mostly focused on preventing catastrophic forgetting, as opposed to improving things over time. And furthermore, it is more focused on facts, skills, etc., as opposed to the quality of the underlying representations.

So putting all this together, and considering the fact that creativity is only one of many barriers to humanlike AI, I do think there is still a very long way to go. That said, I want to express some humility here. These are all just hypotheses. I think we need empirical verification.


Can we empirically test AI ability to do humanlike AI research?
Ongoing work: evaluating agents at long-horizon AI research.
Many similar previous efforts, but we take care to avoid:
• Narrow problems with verifiable ground truth
• Human-in-the-loop without clear boundaries
• Scaffold overfit to the research question
https://cruxevals.com/research-notes/long-horizon-ai-rd-evals
35

Speaking of empirical measurement, we have an ongoing project testing AI's ability to do humanlike AI research.

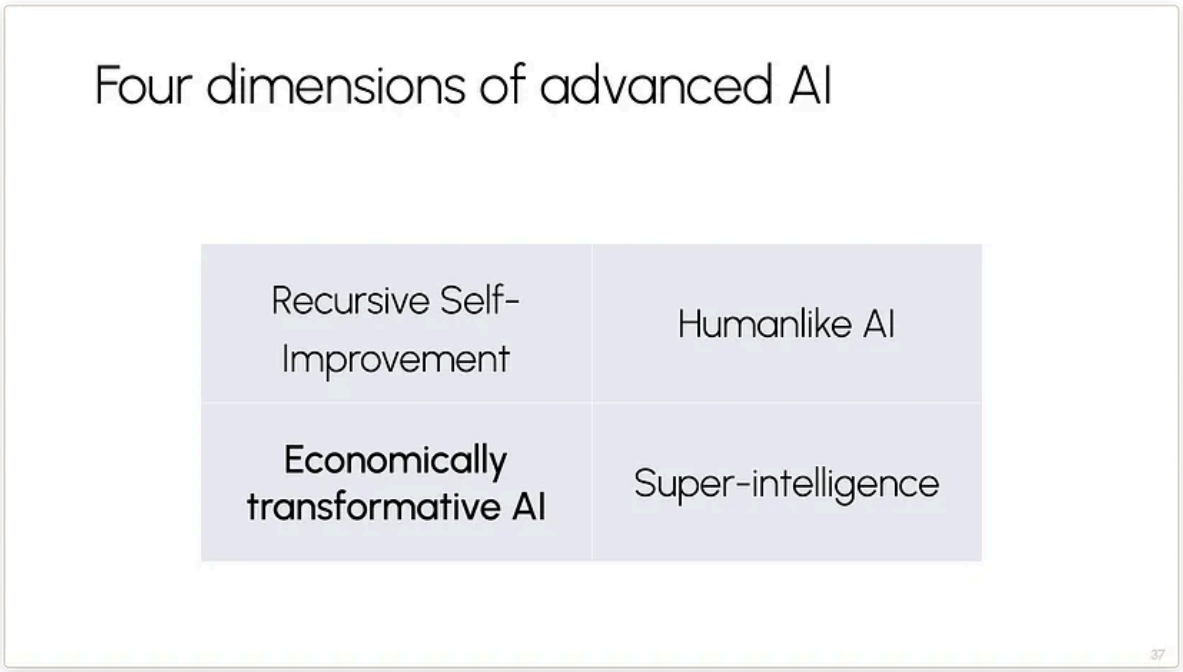
We give AI agents a budget of a few thousand dollars and a machine learning research problem — a problem that researchers have already worked on and written a paper about, but not yet made public on arXiv. This combines two properties: it gives us a team of human judges who've thought deeply about this problem for months and can therefore judge the AI output, but their thinking is not yet online, so it prevents AI cheating or contamination.

Many others have done auto-research experiments, but we're trying to do some things differently — in particular, picking somewhat open-ended problems, so that we can actually test the ability of AI agents at exercising judgment and creativity. We hope to release detailed findings very soon.

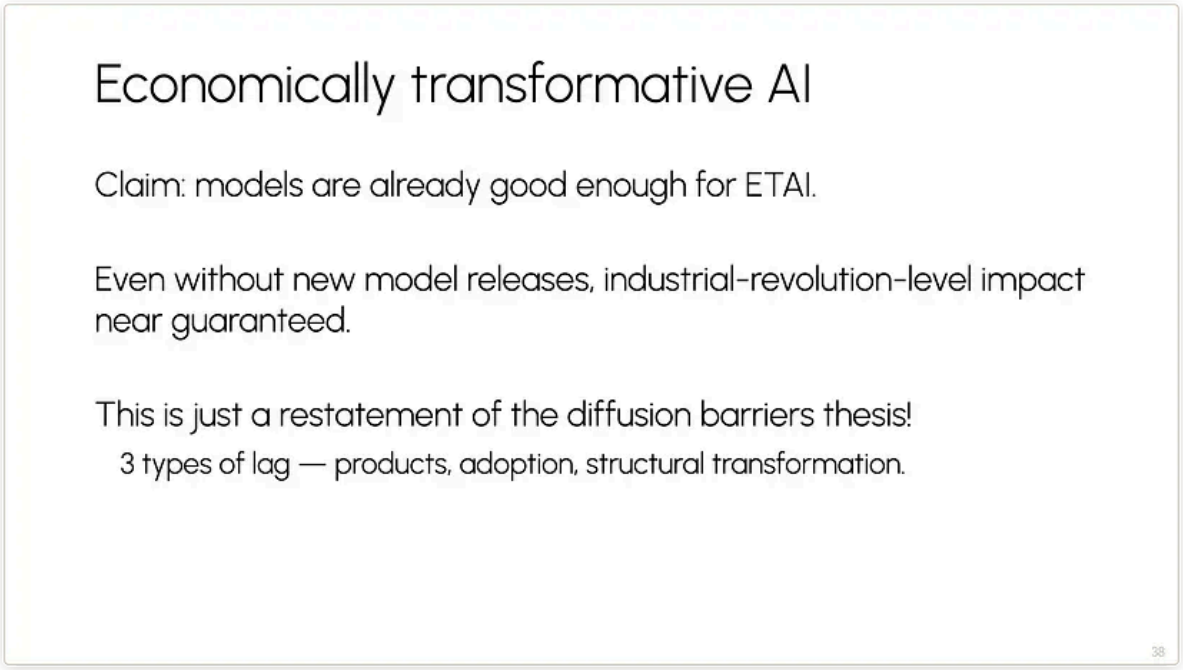
We call this "open-world evaluation"
Open-World Evaluations
for Measuring Frontier AI Capabilities
Sayash Kapoor[1*] Peter Kirgis[1] Andrew Schwartz[1,2] Stephan Rabanser[1]
J.J. Allaire[3] Rishi Bommasani[4] Harry Coppock[5] Magda Dubois[5] Gillian Hadfield[6]
Andy Hall[4] Sara Hooker[7] Seth Lazar[8,9] Steve Newman[9]
Dimitris Papailiopoulos[10,11] Shoshannah Tekofsky[12] Helen Toner[13] Cozmin Ududec[5]
Arvind Narayanan[1]
[1]Princeton University [2]Cornflower Labs [3]Meridian Labs [4]Stanford University
[5]UK AI Security Institute [6]Johns Hopkins University [7]Adaption Labs
[8]Australian National University [9]Golden Gate Institute for AI [10]UW Madison
[11]Microsoft Research [12]AI Digest [13]Georgetown University (CSET)
Interested in leading future open-world evals?
See https://cruxevals.com
36

This builds on a foundation that we call open-world evaluation. We've completed one open-world evaluation of getting agents to build and upload an app to the Apple App Store — that's not about recursive self-improvement, but it tests a different kind of thing. We have assembled a great team, with people from many universities, a couple of companies, as well as the UK AI Security Institute.

And by the way, we are hiring a researcher to lead some of our future open-world evaluations. If you're interested, go check out our website.



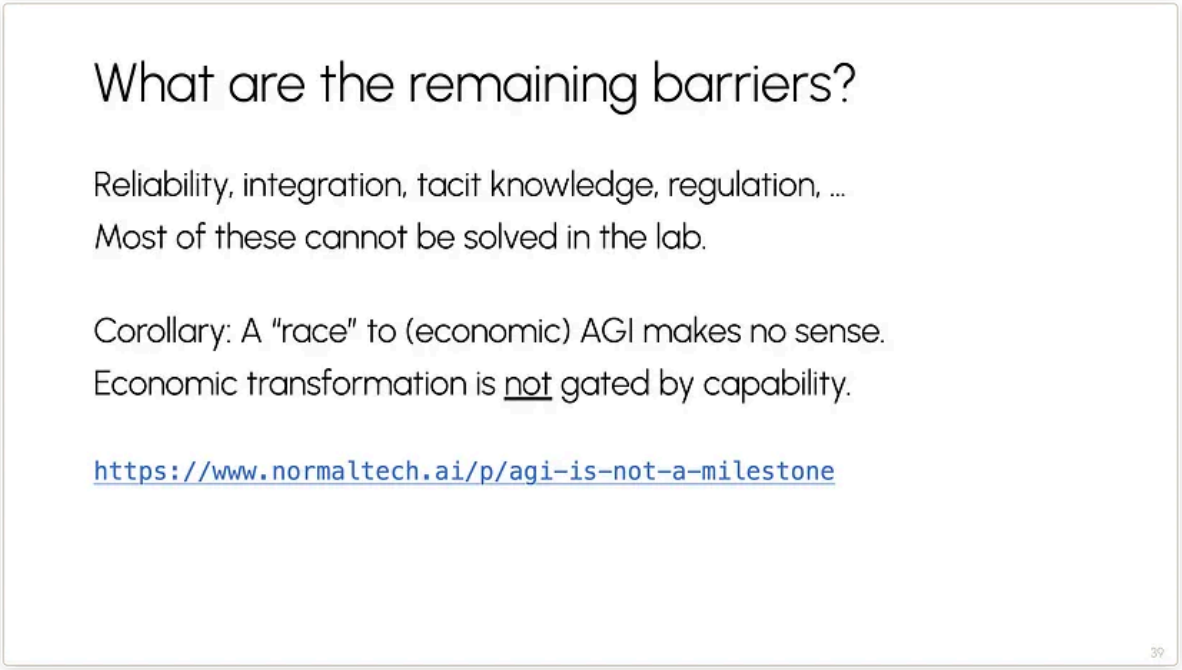
Now let's turn to the third of the four dimensions.


Economically transformative AI
Claim: models are already good enough for ETAI.
Even without new model releases, industrial-revolution-level impact near guaranteed.
This is just a restatement of the diffusion barriers thesis!
3 types of lag — products, adoption, structural transformation.
38

Some people say that AGI is already here. Well, this is one way to interpret that claim, and I happen to agree with it.

This might seem surprising given the skepticism of rapid economic impacts that I expressed in Part 1. But this is actually not only consistent, but in fact a restatement of Part 1. My point there was that the barriers are downstream, and therefore model improvements won't rapidly change the economy. But by the very same token, because the barriers are downstream, even without model improvements, those

barriers are gradually going to get addressed. Yes, it might take a couple of decades, but we will definitely get there.

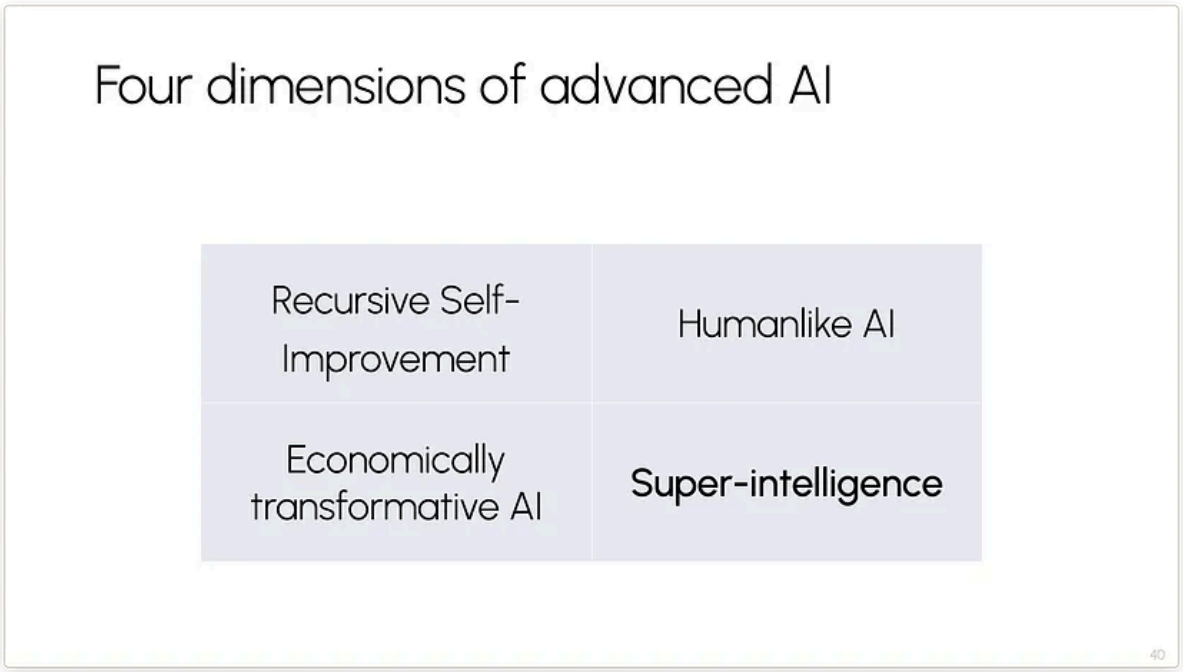

What are the remaining barriers?
Reliability, integration, tacit knowledge, regulation, ...
Most of these cannot be solved in the lab.
Corollary: A "race" to (economic) AGI makes no sense.
Economic transformation is not gated by capability.
https://www.normaltech.ai/p/agi-is-not-a-milestone
39

There are many barriers: reliability (already discussed); integrating AI models into various existing systems; tacit knowledge from domains like medicine or law or various other professions that need to be made available to the models; regulation that often straight up prohibits today's AI systems from being used in productive ways — in many cases there are good reasons for that regulation, but it will need to be modified in order to be able to enable adoption.

The key point is that these are not things that will be solved in the lab. These are not things that are going to be addressed by the next model release on Tuesday. They will be addressed gradually, through gradual adoption, over a period of decades.

From a geopolitics or strategic perspective, some people advocate a race to AGI, arguing that, similar to a Manhattan Project, the country that gets to some capability milestone first is the one that is going to reap economic rewards. I strongly disagree.

There is no particular capability milestone that will unlock all of this economic potential. The economic potential is already there. It really depends on all these downstream actions that we take. It is not gated by capability.



Okay. Now let me talk about the last of these dimensions, namely superintelligence.

Three claims about superintelligence
1. RSI won't get us there
• Thought experiment 1: curing cancer (external barriers)
• Thought experiment 2: weather prediction (inherent limits)
2. Human intelligence is limited by learning & tools, not biology.
• Thought experiment 3: time travel from 200,000 years ago
• Consequence: AI progress also improves human intelligence. We are the superintelligences.
3. Worries about superintelligence usually minimize human agency.
• Thought experiment 4: AI owning & running a company.
• Blunt view: If you treat such outcomes as inevitable and pin your hopes on "aligning" models, you're against safety, not for it.
41

I gave the example of medical trials earlier. Another example: Do you think that we will get superintelligence that can predict the weather precisely a year into the future? We know that that is basically a mathematical impossibility because of the theory of chaos in nonlinear dynamical systems. Our position is that a surprising number of tasks are similar to weather prediction, in that there are inherent limits and we are pretty much already at those limits.

Even for tasks where we're not at that limit, the idea that there is going to be AI superintelligence that's going to obviate humans relies on a fundamental

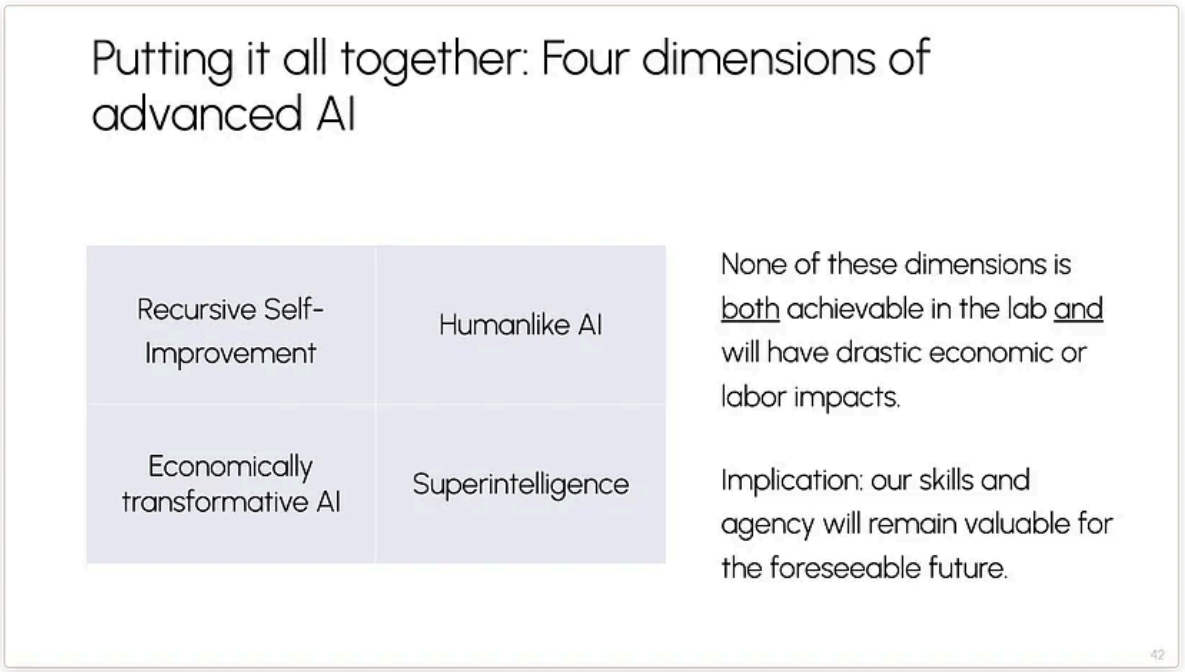
misunderstanding of human intelligence. I claim that in the vast majority of tasks, our performance is not limited by our biology — it is rather limited by our learning and tools.

If you imagine someone from the ancient past time traveling to our world, we're superintelligent compared to that person. And it's not because our biology is better, but because they don't have the benefit of all the learning that we've been through and all the tools, especially digital tools, that we're able to use in order to be productive at whatever it is that we do.

And AI is one such tool. So what that means in our framework is that improvements in AI are actually improving human intelligence, not just AI intelligence. So we have a race between the performance of AI-augmented humans on the one hand and the performance of AI systems acting alone on the other hand. I think we can ensure that we are the superintelligences of the future and we win that race, as opposed to allowing AI systems to act alone in ways that would threaten our future and control.

Many people are pessimistic about this. They bring up thought experiments such as AI running and owning companies in the future. Their view is that if the AI system is not aligned, terrible things might happen — the AI could turn into a paperclip maximizer or do other catastrophic stuff. Our perspective is very different. If you're imagining a future where AI is actually owning and running companies and hiring and firing people, that's already dystopian. It doesn't matter if that AI is aligned or not. The consequences for human dignity. democratic governance, etc. are already catastrophic.

So if your view on safety is to treat this as an inevitable future and just to hope for "alignment" to solve that problem — pardon me for being a little bit blunt here — it feels to me that you're against safety, not for safety. I think it will take a lot of hard work to ensure that we don't irresponsibly deploy AI systems in this kind of fashion. But I do think we can get there. And that's the reason I've spent a lot of my career advising policymakers. We're going to need policy and we're going to need politics, and it's going to be hard, but let's not give up that fight before it even starts.


Putting it all together: Four dimensions of advanced AI
Recursive Self-Improvement
Humanlike AI
Economically transformative AI
Superintelligence
None of these dimensions is both achievable in the lab and will have drastic economic or labor impacts.
Implication: our skills and agency will remain valuable for the foreseeable future.
42

For example, RSI is achievable in the lab, but it won't immediately put people out of work. On the other hand, economically transformative AI is going to happen, but it's not because of the next model release — it's because of things that will gradually happen over the next couple of decades.

Connecting back to the title of this talk: there's no world in which something that an AI company will decide to do in a lab will put us all out of work. Yes, there are risks to be worried about. Yes, things are going to change. But we have agency over how AI gets deployed, and that process will unfold over decades. Again, this is not guaranteed, but this is the future that I want to work towards, and this is the reason why I'm cautiously optimistic.

Outline of this talk

1. The "normal technology" framework is correct and useful unless/until there's a discontinuity

2. Why I'm not losing sleep over Recursive Self-Improvement

3. How human roles shift in a world with advanced AI

43

Let me take the last fifteen minutes to talk about the flip side. I do think a lot of things are going to change. What are some of them?

General observation: Purely technical skills get devalued

This is a long-running trend in programming.

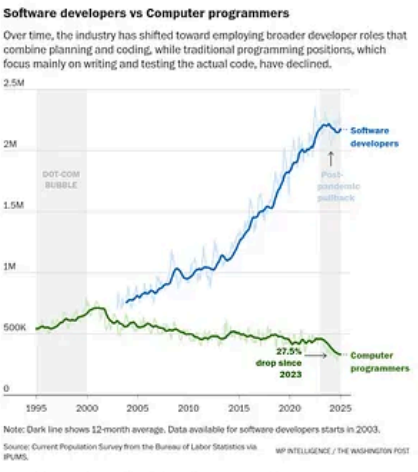

Software developers vs Computer programmers
Over time, the industry has shifted toward employing broader developer roles that combine planning and coding, while traditional programming positions, which focus mainly on writing and testing the actual code, have declined.
2.5M
2M
1.5M
1M
500K
0
DOT-COM BUBBLE
Post-pandemic pullback
Software developers
27.5% drop since 2023
Computer programmers
1995
2000
2005
2010
2015
2020
2025
Note: Dark line shows 12-month average. Data available for software developers starts in 2003.
Source: Current Population Survey from the Bureau of Labor Statistics via IPUMS.
WP INTELLIGENCE / THE WASHINGTON POST

Chart by the Washington Post

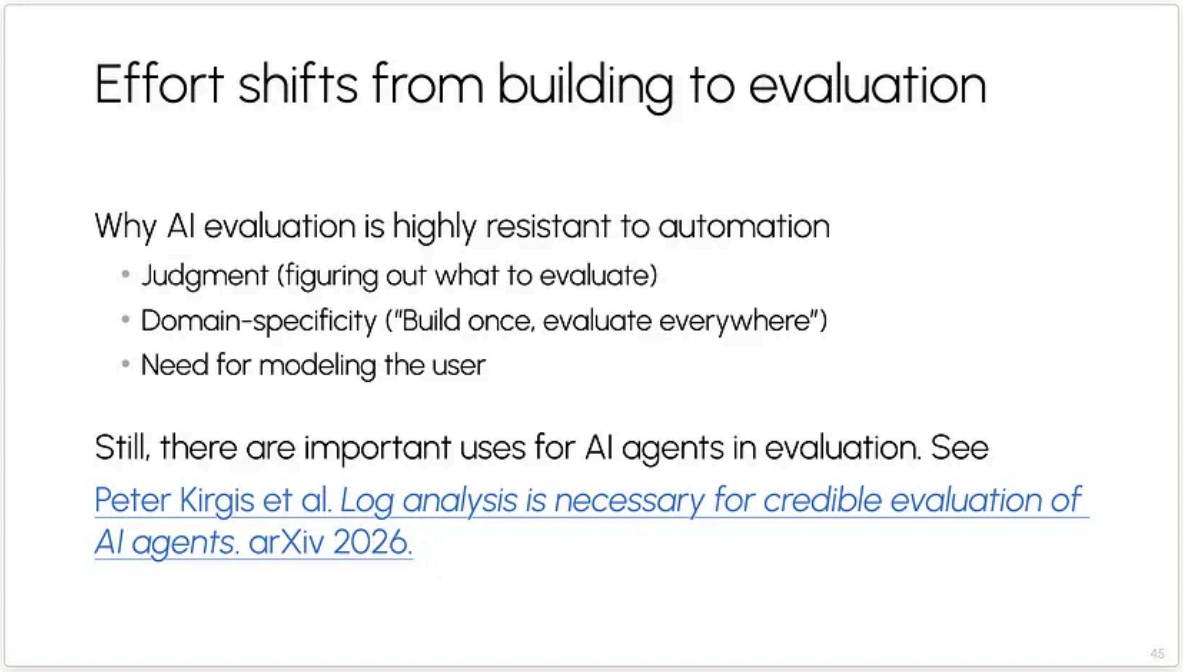
44

Technical skills tend to be verifiable tasks, and AI will continue to get better at them.

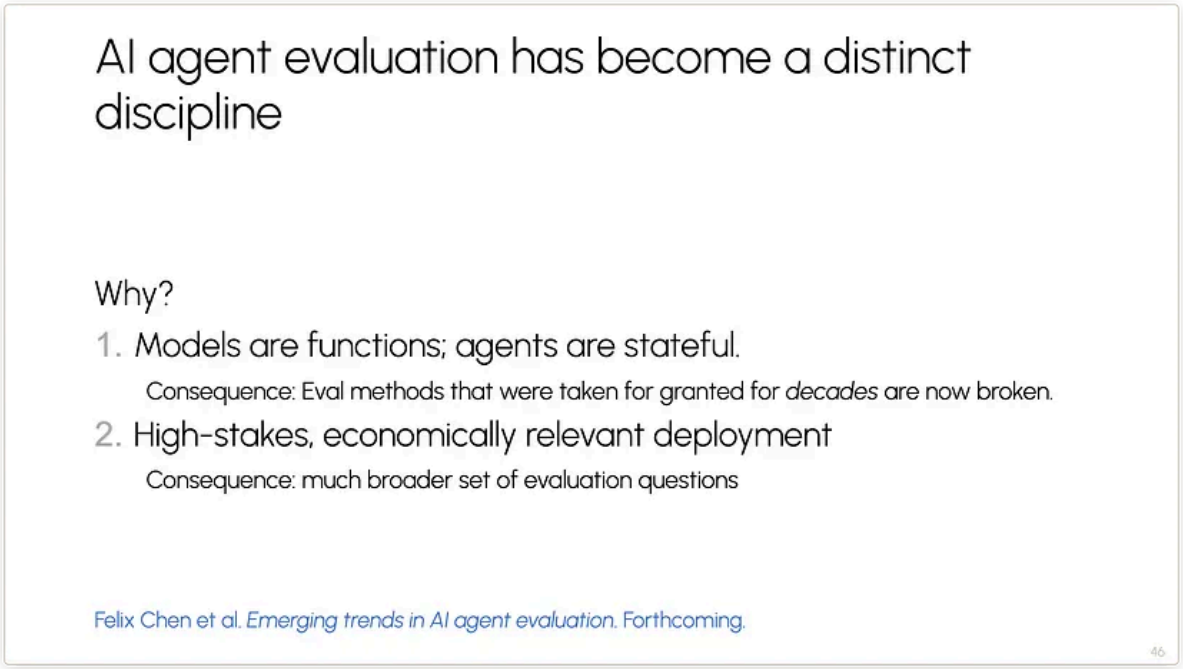
Over twenty years ago, there started to be a stark difference in the labor demand for programming jobs versus software engineering jobs. Programming jobs are conceived narrowly around the technical skills of coding and debugging. Software engineering jobs are responsible for all three layers of the decide, execute, deliver sandwich that I talked about — figuring out what even needs to be built,

understanding customers, that sort of thing. It requires domain knowledge, judgment, and more.

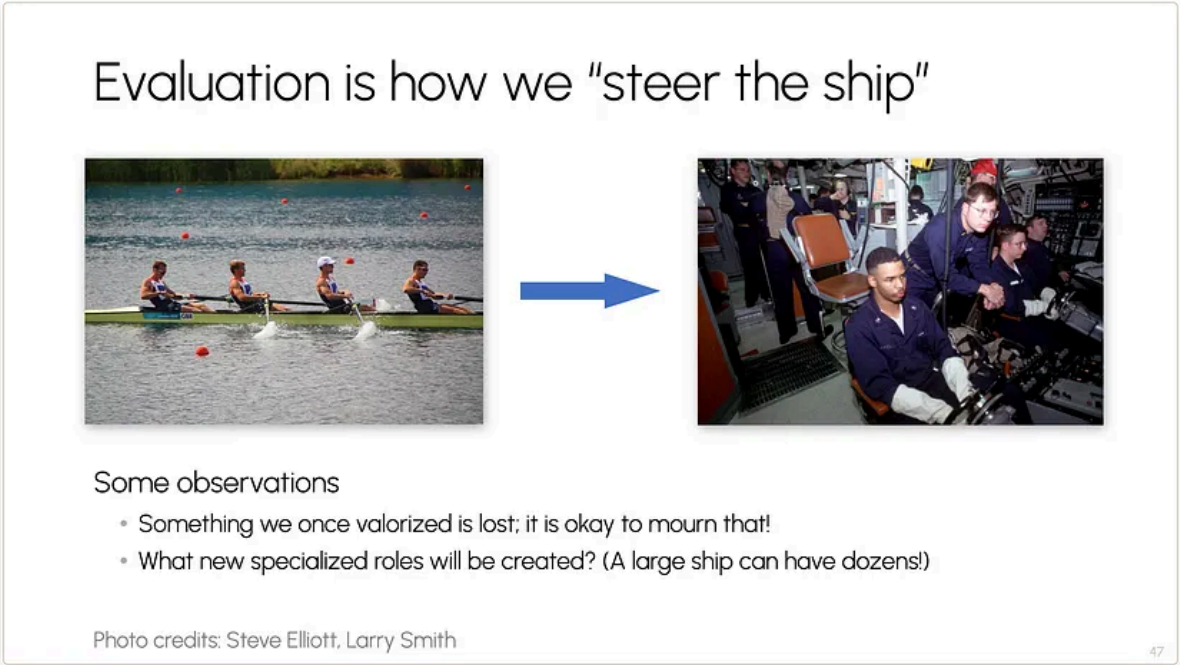
I predict that this will happen in more and more fields over time.


Effort shifts from building to evaluation
Why AI evaluation is highly resistant to automation
• Judgment (figuring out what to evaluate)
• Domain-specificity ("Build once, evaluate everywhere")
• Need for modeling the user
Still, there are important uses for AI agents in evaluation. See
Peter Kirgis et al. Log analysis is necessary for credible evaluation of AI agents. arXiv 2026.
45

A recurring pattern I've observed — effort shifts from building systems to evaluating systems. As I've mentioned, I lead a team working on AI agent evaluation. LLMs and agents are general purpose. So each time capability goes up, it creates demand for evaluation in a legal setting or a journalistic setting or whatever other setting, and that's not work that is scalable.


AI agent evaluation has become a distinct discipline
Why?
1. Models are functions; agents are stateful.
Consequence: Eval methods that were taken for granted for decades are now broken.
2. High-stakes, economically relevant deployment
Consequence: much broader set of evaluation questions
Felix Chen et al. Emerging trends in AI agent evaluation. Forthcoming.
46

Not only is AI agent evaluation resistant to automation — it has become sufficiently specialized that the set of people and teams working on evaluation is starting to diverge from the set of people and teams building and pushing the state of the art in AI agents. This new community is developing a new set of best practices around what it means to rigorously evaluate agents, and we have a forthcoming paper that is going to look at that in some detail.

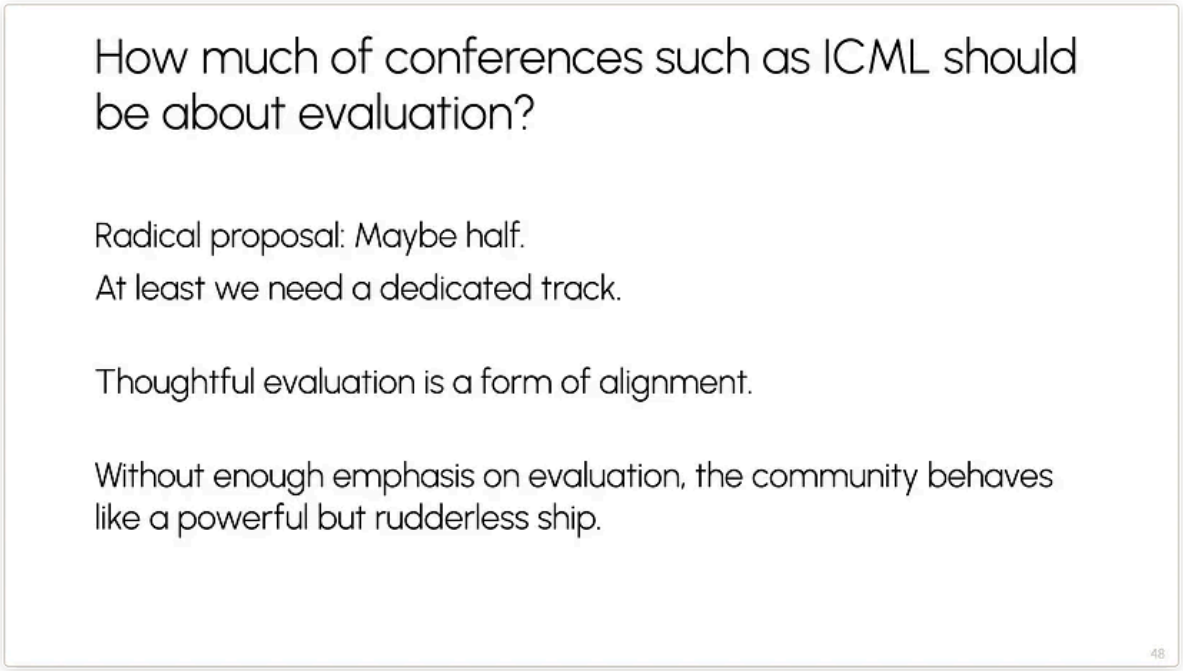

Evaluation is how we "steer the ship"
Some observations
• Something we once valorized is lost; it is okay to mourn that!
• What new specialized roles will be created? (A large ship can have dozens!)
Photo credits: Steve Elliott, Larry Smith
47

Here's an a metaphor to help explain the shift I see in our community. Imagine that in the past most boats were rowboats, and the work of the humans was in physically moving the boat. There was no separate specialized role around steering the boat: when you're rowing the boat, you're also figuring out which way to row the boat.

But what happened when the physical work of moving of the ship could be delegated to the engines? The human jobs didn't go away. In fact, they became much more specialized. Modern ships have very complicated control panels, and they might have dozens of different specialized roles that are focused on where the ship should go and how it should get there.

I would argue that we're seeing a similar shift in AI/ML. In the past, most of our work was on building — we didn't need separate roles for evaluation. That has changed now. We're still in the early stages of this process, but I think over time more and more of the building, because it's a verifiable task, will be able to be done by AI, whereas it's the evaluation — figuring out where we should go as a community, figuring out what are the desirable properties in AI systems — that is very resistant to automation. So a greater fraction of the community's attention will have to focus on evaluation over time, compared to where it is today.

In rowing, physical strength was valorized, and today we might be sad about the fact that sailors don't need to be strong and it's all a bunch of nerds. Similarly, today in the AI community, there's still great value in deep technical understanding of models and systems, and that is considered the coolest thing; the most important thing; that's the thing that commands a lot of value. But I think in the future that will become less important than exercising judgment and all of these fuzzy things that have a lower status in the AI community today. That's a mental shift that we're maybe not quite prepared for. Many of us will be sad about it, and that's okay.


How much of conferences such as ICML should be about evaluation?
Radical proposal: Maybe half.
At least we need a dedicated track.
Thoughtful evaluation is a form of alignment.
Without enough emphasis on evaluation, the community behaves like a powerful but rudderless ship.
48

One practical consequence of this shift: consider a conference like this one. What fraction of the conference should be dedicated to evaluation papers?

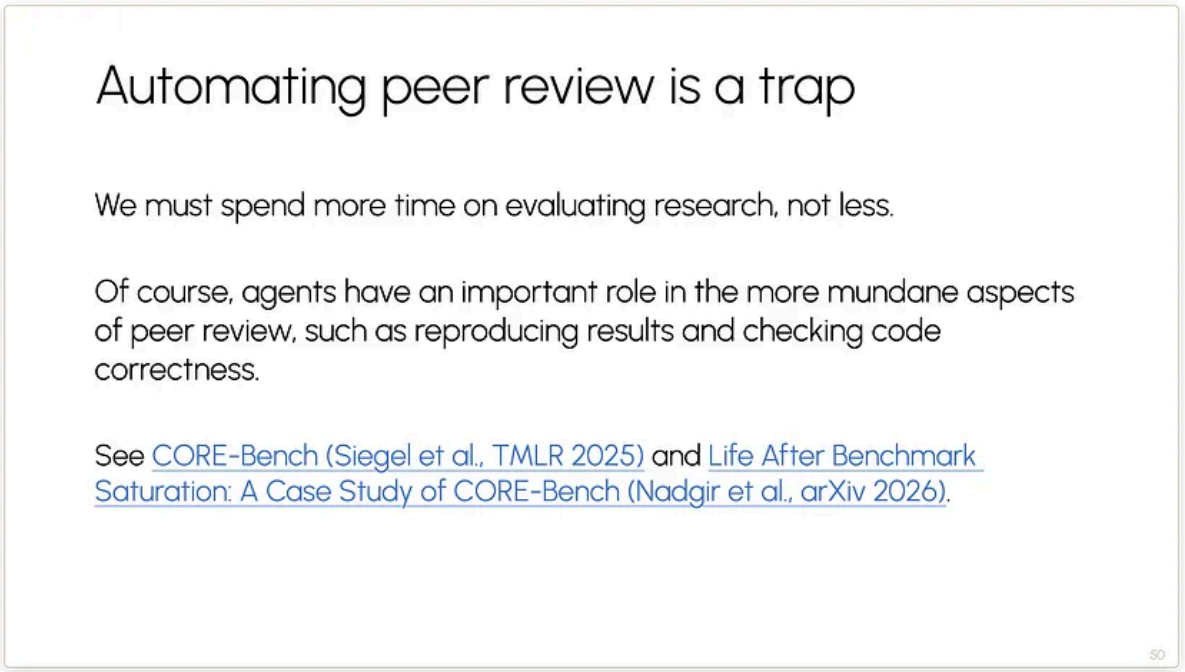
I don't know, but I think maybe a lot more than it is today. Maybe around half — I'm just throwing a number out there — which is orders of magnitude more than it is today. At the very least we need a dedicated track. We don't have a dedicated track. NeurIPS does have one, and it has been growing in popularity over time, and I think that is a good thing.

And I would go further and argue that thoughtful evaluation of AI systems is a form of alignment. It's not aligning the AI system itself, but it's aligning the community as a whole — thinking about where we're currently going and where we want to go, and trying to align those two to each other. And without enough emphasis on evaluation, my worry is that the community — going back to the ship metaphor — will behave like a rudderless ship: very powerful, but where we don't collectively have control over what kind of AI we want to develop.

How we evaluate research itself must change
Position: There are futures that benchmark-driven AI cannot see.
Sobhan Lotfi, Ava Iranmanesh et al. ICML 2026 Oral.
Benchmark-based evaluation is efficient but greatly narrows the field.
49

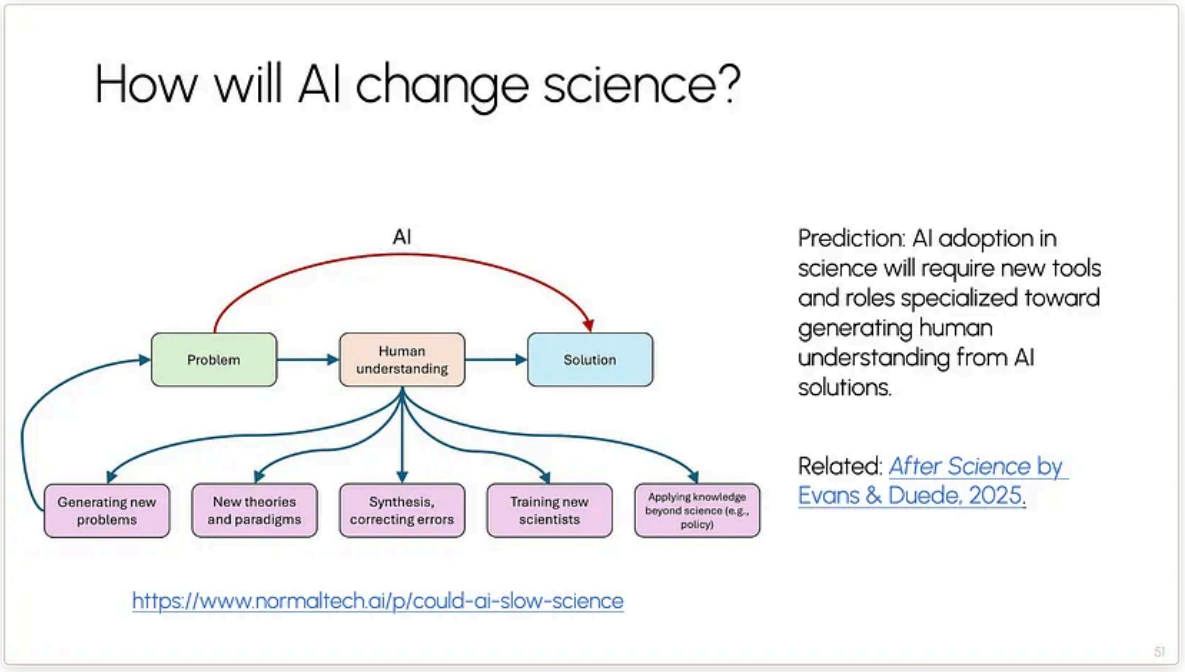
I had a small role in this cool position paper. A one-sentence version of the argument: we need to move beyond benchmarks being the be-all and end-all of what AI research is considered valuable.

What benchmarks give us is efficiency. We don't have to spend years peer-reviewing papers — if it beats the state of the art, we know it's probably a good paper. But unfortunately, it has narrowed the collective vision of the community. We're searching under streetlights — we are searching the kinds of things that are easy to search in a benchmark-driven evaluation regime. We need to move beyond that. That means evaluating how good a paper is will become much more costly. Sadly, that is a cost that we will have to pay. Right now the community is trying to rush in the opposite direction.


Automating peer review is a trap
We must spend more time on evaluating research, not less.
Of course, agents have an important role in the more mundane aspects of peer review, such as reproducing results and checking code correctness.
See CORE-Bench (Siegel et al., TMLR 2025) and Life After Benchmark Saturation: A Case Study of CORE-Bench (Nadgir et al., arXiv 2026).
50

There is a big temptation towards automating peer review. And with all due respect, I think this is a trap. If we automate peer review, we essentially give up control over the direction of progress to AI systems themselves. That seems like a fundamental misallocation of effort.

If we automate the mundane parts, human time can be freed up to think more deeply about the parts that require judgment.

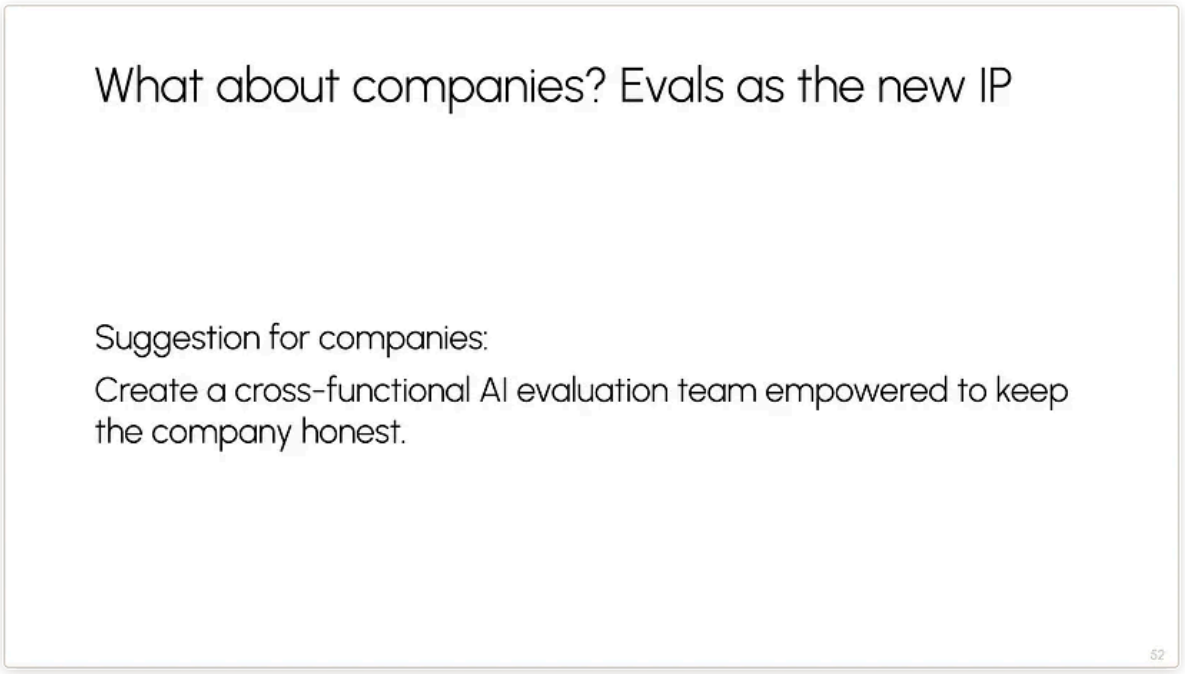

How will AI change science?
AI
Problem
Human understanding
Solution
Generating new problems
New theories and paradigms
Synthesis, correcting errors
Training new scientists
Applying knowledge beyond science (e.g., policy)
Prediction: AI adoption in science will require new tools and roles specialized toward generating human understanding from AI solutions.
Related: After Science by Evans & Duede, 2025.
https://www.normaltech.ai/p/could-ai-slow-science
51

Generalizing from AI research to scientific research as a whole, we have an essay that argues that visions of “automating science” rely on a fundamental misunderstanding — as if the point of science is mere problem solving, and that if we

can use AI to get directly from the problem to the solution, we will have automated and accelerated science.

In our view, human understanding is not some friction to be automated away. It is rather essential. It is central to the very purposes behind why we do science. And if we lose human understanding, we lose all of these things that flow from it.

So my prediction is that if we are going to have increased use of AI agents in science, there will be new tools and new roles that are specialized towards looking at those AI solutions and backing out the human understanding from those solutions. Because it is essential to preserve human understanding.


What about companies? Evals as the new IP
Suggestion for companies:
Create a cross-functional AI evaluation team empowered to keep the company honest.
52

Many people in the corporate world are saying that evals are the new IP. Without going into the details, it is very much the same phenomenon playing out once again — effort shifting from purely building to a mix of building and evaluation of systems.

(Note: in this post I explain the idea of cross-functional eval teams to keep companies from fooling themselves.)

How effort shifts from building to evaluation

• AI agents	New evaluation community
• AI research	Beyond benchmark-driven peer review
• Science, broadly	Understanding & evaluating AI solutions
• Software engineering	Decide-execute-deliver
• Companies	Evals as the new IP

53

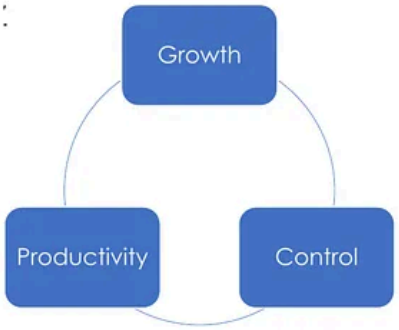
If there's only one thing you take away from this talk, it should probably be this slide. What I've shown throughout the talk is that in many different areas, because the verifiable tasks can be handled by AI, some or much of the effort shifts from building to evaluation.

Effort shifts from "rowing the boat" to "steering the ship, navigating the ship, and figuring out where we even want to go".

I think this is a powerful metaphor that will allow us to predict how human roles will shift over the next decade or two.

Personal reflections on adapting to AI

- The floor (AI capability) keeps moving up but so does the ceiling (AI-assisted human capability) — if we push it.
- The optimal balance between productivity and growth has shifted.
- Embracing "cognitive sweat"
- Staying in control
  - Resisting the black-box temptation
  - Avoiding the dependence spiral



54

In the last few minutes, let me give you some personal reflections on how I've been struggling through this challenge — the fact that AI capabilities are rapidly improving — in my own research workflows, as I'm sure many of you are. I think everyone should choose their own path, but hopefully there are some interesting ideas here for you to consider.

By floor I mean what AI can do on its own. By ceiling I mean what AI allows us to do by augmenting our capabilities — the ability to take on new and ambitious projects that were not possible before AI. But the ceiling is not going up automatically. It is only if we work on pushing the ceiling up.

I find that I'm spending something like 10 hours per week just learning and experimenting with new workflows, as well as learning new topics. So one way to think about it is that AI enables big productivity improvements, and I've been trying to take that time saved and "reinvest" it into long-term growth and picking up new and complementary skills.

I've learned that if I don't feel exhausted at the end of the day, I've done something wrong. I've offloaded too much to AI. I've sacrificed too much of my long-term growth in the pursuit of short-term productivity.

Growth and productivity are two legs of a three-legged stool that we need to learn how to balance. And the third leg of that stool is staying in control.

Here are two heuristics that I've tried to use in order to do that. The first is resisting the black-box temptation. Companies want us to use agents as black boxes — just prompt it and it will go off and do its thing. I think that's a trap. I think that's very dangerous, and it will lead us to gradually give up control over time.

Second, and related, is what I call the dependence spiral. It's very tempting to use AI for tasks that I am myself not yet an expert at, because learning new things is of course very hard. But that leads to my losing whatever little skills I have in that task over. It's much better in the long run if I first put in the time to master the task myself before I use AI to augment my productivity.

The vision: co-superintelligence
We can ensure that the superintelligences of the future are AI-augmented humans rather than AI systems acting alone.
A ~~bicycle~~ crane
for the mind
55

None of this is easy, but if we get this right, the vision is tantalizing and empowering.

Computers have often been called bicycles for the mind. I think AI can be more than that. I think AI can be a crane for the mind, if you will indulge my metaphor, in the sense that it can amplify our potential to previously unimaginable heights. Getting there can seem daunting. It has an incredible learning curve. I feel like I'm on a treadmill all the time, but I'm very excited about it. I think it's a fun challenge, and I think it's worth fighting this fight. Compared to five years ago, in a way, I feel ... maybe superintelligent is not the right word, but I feel like I have superpowers, given the extent to which AI allows me to take on new ambitious things that were not possible before, and push myself harder than was possible before.

And I think we can ensure that this remains the case even as AI capabilities advance, for the foreseeable future. It might be that in some distant future this becomes impossible to do, but it is very premature to give up the fight now. I certainly plan to continue this fight, and I hope to work towards this vision of co-superintelligence, and I hope you will join me. That is my closing thought.

Acknowledgments

Sayash Kapoor

Stephan Rabanser

Peter Kirgis

Nitya Nadgir

Other key contributors and alumni: Rishi Bommasani, Andrew Schwartz, Rob Schwartz, Felix Chen, Kangheng Liu, Saiteja Utpala, Benedikt Stroebl, Veniamin Veselovsky, Zachary Siegel

Takeaways

- The "normal technology" framework
  - Models → products → early adoption → adaptation
  - Predicts that impacts will unfold over decades
- AI agent reliability remains a barrier to automation.
  - Build collaboration agents, not just automation agents.
- As AI advances, human roles paradoxically expand, not shrink.
- No lab milestone will lead to sudden economic/labor impacts.
- As verifiable tasks get automated, work shifts from building to evaluation.
- The vision of co-superintelligence is worth fighting for.